Math 211 Class notes
Spring 2026

To accompany
Essential Statistics: Exploring the World through Data
by *Gould, Wong,* and *Ryan*

Peter Westerbaan

Last updated: January 4, 2026

Table Of Contents

Unit 1 – Data and the Stories They Tell . . . 1

1.1: What Are Data? . . . 1

1.2: Classifying and Storing Data . . . 2

1.4: Organizing Categorical Data . . . 5

1.5: Collecting Data to Understand Causality . . . 6

2.1: Visualizing Variation in Numerical Data . . . 8

2.2: Summarizing Important Features of a Numerical Distribution . . . 12

2.3: Visualizing Variation in Categorical Variables . . . 16

2.4: Summarizing Categorical Distributions . . . 19

2.5: Interpreting Graphs . . . 22

3.1: Summaries for Symmetric Distributions . . . 25

3.2: What's Unusual? The Empirical Rule and z-Scores . . . 29

3.3: Summaries for Skewed Distributions . . . 34

3.4: Comparing Measures of Center . . . 38

3.5: Using Boxplots for Displaying Summaries . . . 41

Unit 2 – Using Models to Describe the World . . . 43

4.1: Visualizing Variability with a Scatterplot . . . 43

4.2: Measuring Strength of Association with Correlation . . . 45

4.3: Modeling Linear Trends . . . 49

4.4: Evaluating the Linear Model . . . 53

6.1: Probability Distributions Are Models of Random Experiments . . . 54

6.2: The Normal Model . . . 58

7.1: Learning about the World through Surveys . . . 64

7.2: Measuring the Quality of a Survey . . . 68

7.3: The Central Limit Theorem for Sample Proportions . . . 71

Unit 3 – Inference: Turning Data into Knowledge 74

7.4: Estimating the Population Proportion with Confidence Intervals 74

8.1: The Essential Ingredients of Hypothesis Testing 77

8.2: Hypothesis Testing in Four Steps 80

8.3: Hypothesis Tests in Detail 83

9.1: Sample Means of Random Samples 85

9.2: The Central Limit Theorem for Sample Means 86

9.3: Answering Questions about the Mean of a Population 90

9.4: Hypothesis Testing for Means 95

Math 211 Formula Sheet

Unit 1

Mean: $\overline{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

Standard Deviation: $s = \sqrt{\dfrac{\sum (x_i - \overline{x})^2}{n-1}}$

Variance: $s^2 = \dfrac{\sum (x_i - \overline{x})^2}{n-1}$

z-Score: $z = \dfrac{x - \overline{x}}{s}$

Interquartile Range: $IQR = Q_3 - Q_1$

Unit 2

Correlation Coefficient: $r = \dfrac{\sum z_x z_y}{n-1}$

Regression line: $y = a + bx$, where $b = r\dfrac{s_y}{s_x}$ and $a = \overline{y} - b\overline{x}$

Standard Error: $SE = \sqrt{\dfrac{p(1-p)}{n}}$

Unit 3

Standard error (proportions) $SE = \sqrt{\dfrac{p(1-p)}{n}}$

Confidence intervals (proportions) $\hat{p} \pm m$ where $m = z^* SE_{\text{est}}$, $SE_{\text{est}} = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

z test statistic (proportions) $z = \dfrac{\hat{p} - p_0}{SE}$ where $SE = \sqrt{\dfrac{p_0(1-p_0)}{n}}$

Standard error (means) $SE = \dfrac{\sigma}{\sqrt{n}}$

Confidence intervals (means) $\overline{x} \pm m$ where $m = t^* SE_{\text{est}}$, $SE_{\text{est}} = \dfrac{s}{\sqrt{n}}$

t test statistic (means) $t = \dfrac{\overline{x} - \mu}{SE_{\text{est}}}$ where $SE_{\text{est}} = \dfrac{s}{\sqrt{n}}$

1.1: What Are Data?

Statistics rests on two major concepts:

a) Data

b) Variation

Statistics is the science of:

- Collecting
- Organizing
- Summarizing
- Analyzing Data

For the purpose of:

- Answering questions and/or
- Drawing conclusions

Context is important! Some questions you can ask:

- Who, or what, was observed?
- How were they measured?
- Who collected the data?
- Where/when/why were the data collected?
- What variables were measured?
- What are the units of measurement?
- How did they collect the data?

1.2: Classifying and Storing Data

- The collection of data is called a **data set** or a **sample**. The **population** refers to the set or group that contains everything relevant to the data.
- When we collect data, the characteristics of that data (e.g. gender, weight, temperature) are called **variables**.
- Variables can be categorized into two groups:
  - Numerical variables
  - Categorical variables

Example. The following table contains data crash-test dummy studies.

- How many variables does this table have?
- How many observations does this table have?
- For each variable, identify whether it is numerical or categorical:

Make	Model	Doors	Weight	Head Injury
Acura	Integra	2	2350	599
Chevrolet	Camaro	2	3070	733
Chevrolet	S-10 Blazer 4X4	2	3518	834
Ford	Escort	2	2280	551
Ford	Taurus	4	2390	480
Hyundai	Excel	4	2200	757
Mazda	626	4	2590	846
Volkswagen	Passat	4	2990	1182
Toyota	Tercel	4	2120	1138

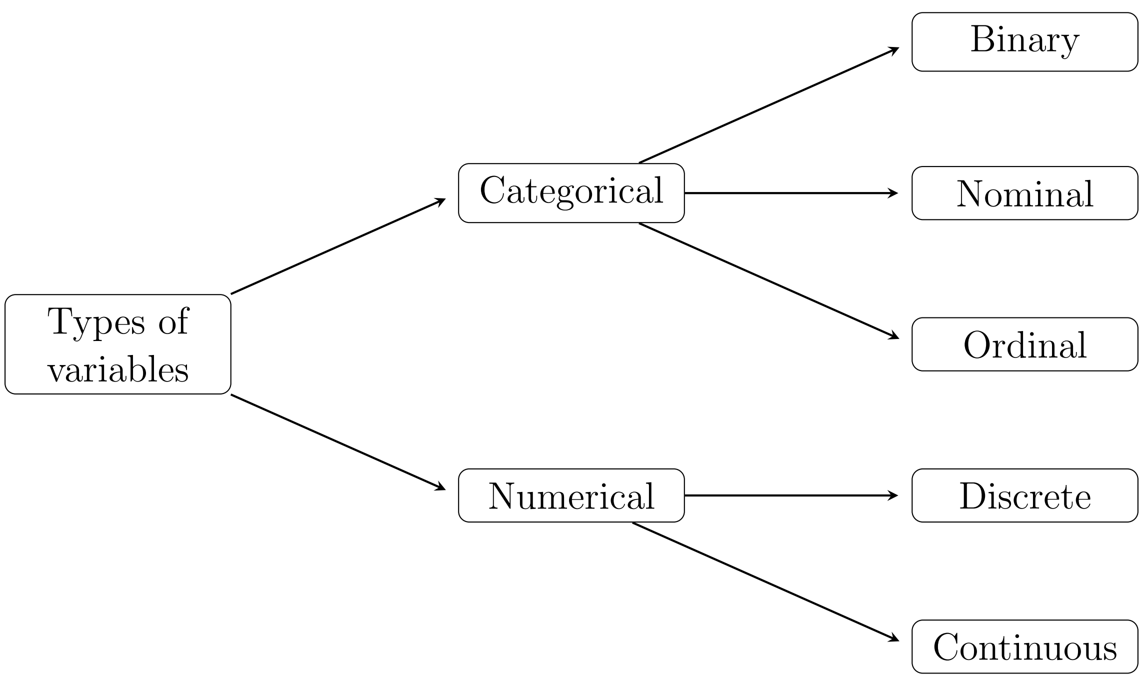
Coding categorical data using numbers:

Weight	Gender	Smoke
7.69	Female	No
0.88	Male	Yes
6.00	Female	No
7.19	Female	No
8.06	Female	No
7.94	Female	No

$\longrightarrow$

Weight	Female	Smoke
7.69	1	0
0.88	0	1
6.00	1	0
7.19	1	0
8.06	1	0
7.94	1	0

We can further break down variables into five types:



Example. Suppose a local store was interested in whether a new product would sell or not. The manager decided to take a random sample of 100 customers over a two-week period and asked each person whether they would buy the product or not and how many times would they buy the product over a six month period.

a) What is the population?

b) What is the sample?

c) What are the variables?

d) Classify each variable as numerical or categorical.

1.4: Organizing Categorical Data

Definition.
In the context of statistics, **frequency** is the number of times a value of a variable is observed in a data set.
Relative frequency (proportion) is a ratio of the frequency of a variable to the total frequency of the group desired. This can be left as a fraction, decimal, or percentage.

Example. The following **two-way table** contains the results of a national survey that asks American youths whether they wear a seat belt while driving or riding in a car:

	Male	Female	Total
Not Always	2	3	
Always	4	8	
Total			

a) Find the total number of males, females, and total participants in this survey.

b) Identify the frequencies, and compute the percentages below:

	Male	Female	Total
Not Always			
Always			
Total			100%

c) Are males or females more likely to take the risk of not wearing a seat belt?

d) Should we use the frequencies or the relative frequencies to make comparisons?

1.5: Collecting Data to Understand Causality

Definition.

- In an **observational study**, we observe individuals and measure variables of interest but do not attempt to influence the responses. (Observe but do not disturb)
- In a **controlled experiment**, we deliberately impose some treatment on (that is, do something to) individuals in order to observe their responses. Researchers assign subjects to a treatment group or control group.
- **Anecdotal evidence** is a story based on someone's experience.

- In an **observational study**, the researcher observes values of the response variable for the sampled subjects, without anything being done to the subjects (such as imposing a treatment).
- In short, an *observational study* merely observes rather than experiments with the study subjects.

Note: Anecdotal evidence and observational studies:

- NEVER point to causality (cause-and effect).
- Only point to an association between variables!

To establish cause-and-effect: Use a controlled experiment!

Definition.
Differences between two groups that could explain different experiences/outcomes are called **confounding variables** or **confounding factors**.

How to design a good experiment ("Gold standard" in experiments):

- Random allocation – participants randomly allocated to treatment and control group
- Use of a placebo if appropriate
  - A **placebo** is a fake treatment (e.g. sugar pill).
  - The **Placebo-Effect** is reacting to a treatment you haven't received.
- Blinding the study – used to avoid bias
  - Single blind – Researcher is unaware of treatment group
  - Double blind – Researcher and subjects are both unaware of treatment group
- Large sample size – accounts for variability

2.1: Visualizing Variation in Numerical Data

Definition.
The **distribution of a sample** of data is a way of organizing the data by recording the

- values that were observed, and
- the frequencies of these values.

Example. Below are the number of goals scored by first year NCAA female soccer players in Division III in the 2016-17 season:

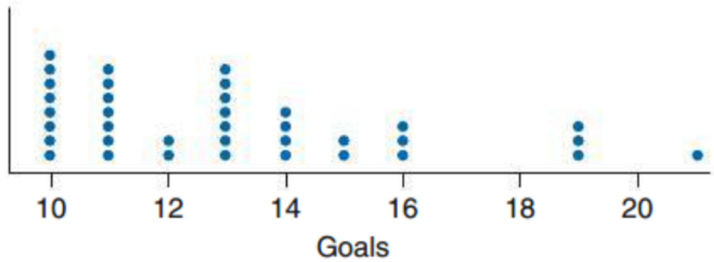
$$11, 14, 16, 13, 13, 10, 13, 11, 16, 21, 13, 19, 10, 10, 14, 13, 10, 13,$$

$$15, 10, 15, 13, 11, 19, 11, 11, 16, 10, 12, 11, 14, 11, 10, 14, 10, 19, 12$$

The **distribution** lists the values *and* the frequencies:

Value	Frequency
10	8
11	7
12	2
13	7
14	4
15	2
16	3
17	0
18	0
19	3
20	0
21	1

A **dotplot** represents the data by using a dot where each value occurs:

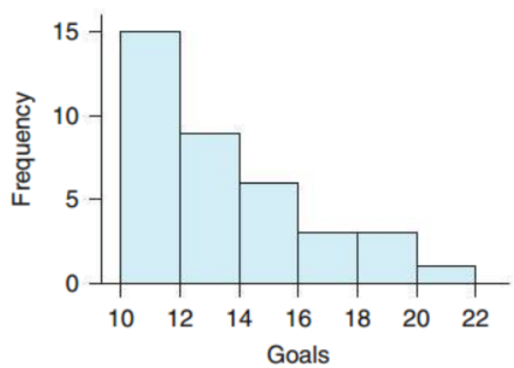


▲ **FIGURE 2.2** Dotplot of the number of goals scored by first-year women soccer players in NCAA Division III, 2016–17. Each dot represents a soccer player. Note that the horizontal axis begins at 10.

Histograms:

A **histogram** represents the data by using bars to indicate how much data lies in each *bin* (also called *interval* or *class*):

► **FIGURE 2.3** Histogram of number of goals for female first-year soccer players in NCAA Division III, 2016–17. The first bar, for example, tells us that 15 players scored between 10 and 12 goals during the season.

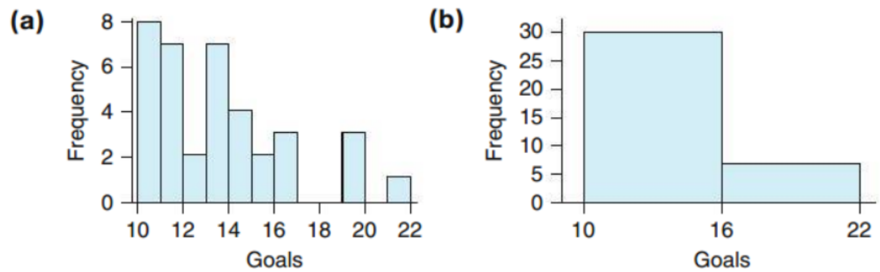


Q: Where do we place data points that lie on a boundary?

Note: Bin size plays a significant role in how the data is represented in a histogram. A bin width that is:

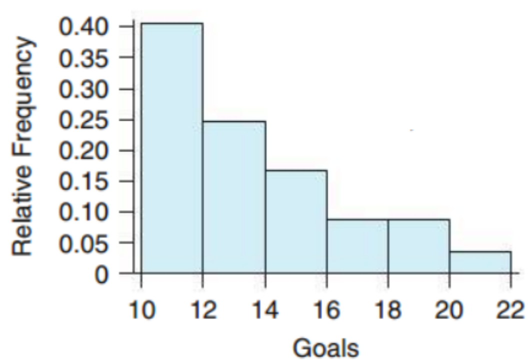
- too narrow shows too much detail.
- too wide hides detail.

▶ **FIGURE 2.4** Two more histograms of goals scored in one season, the same data as in Figure 2.3. **(a)** This histogram has narrow bins and is spiky. **(b)** This histogram has wide bins and offers less detail.



A **relative frequency histogram** changes the units on the vertical axis to represent relative frequencies:

▶ **FIGURE 2.5** Relative frequency histogram of goals scored by first-year women soccer players in NCAA Division III, 2016–17.



Stemplots:

Definition.
A **stemplot** divides each observation into a *stem* and *leaf*. The **leaf** is the last digit in the observation, and the **stem** contains all the digits preceding the leaf.

Example. A collection of college students who said that they drink alcohol were asked how many alcoholic drinks they had consumed in the last seven days. Their answers were:

$$1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 8, 10, 10, 15, 17, 20, 25, 30, 30, 40$$

Stem	Leaves
0	111112223333345556668
1	0057
2	05
3	00
4	0

Example. Below is a stemplot for exam grades. How many grades are between 40% and 59%?

Stem	Leaves
3	8
4	
5	
6	0257
7	00145559
8	0023
9	0025568
10	00

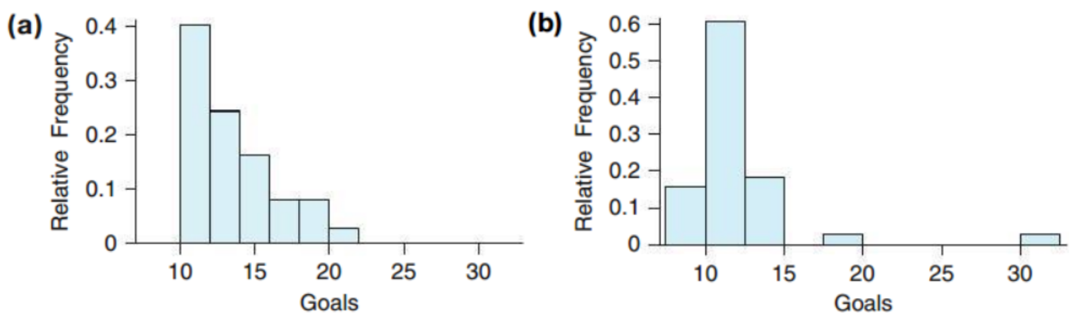
2.2: Summarizing Important Features of a Numerical Distribution

Definition.
When examining a distribution:

- the **center** represents the typical or most common values, and
- the **spread** represents the variability in the data.

Example. Below are the histograms containing the number of goals scored by first year NCAA female (left) and male (right) soccer players in Division III in the 2016-17 season:

▶ **FIGURE 2.9** Distributions of the goals scored for **(a)** first-year women and **(b)** first-year men in Division III soccer in 2017.



– Are their any notable differences in the shapes?

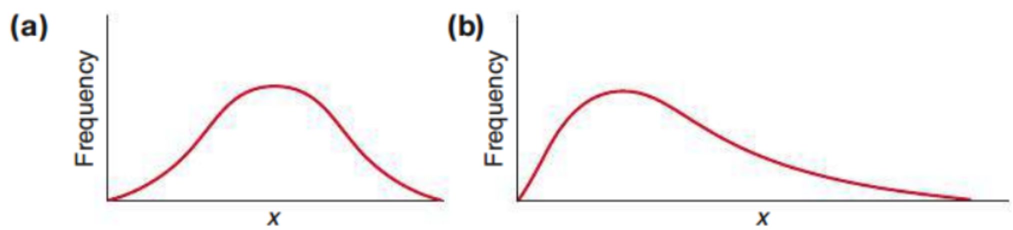
– What is the approximate center for each distribution?

– How do the spreads compare?

Three basic characteristics to consider when examining a distribution's shape:

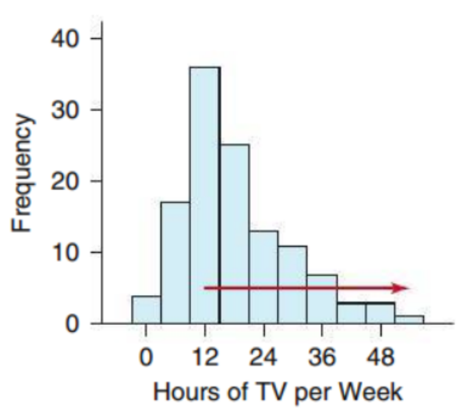
1. Is the distribution symmetric or skewed?
2. How many "mounds" appear?
3. Are unusually large or small values present?

▶ **FIGURE 2.10** Sketches of **(a)** a symmetric distribution and **(b)** a right-skewed distribution.

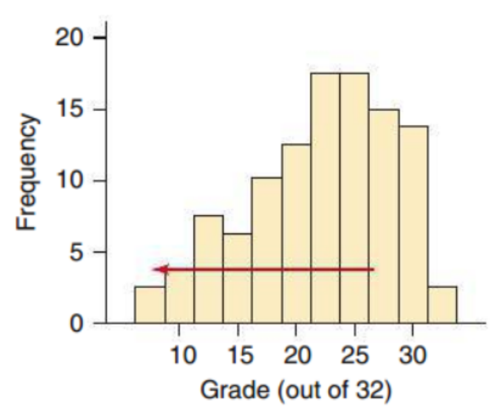


Definition.

- A **right-skewed distribution** has a "tail" that extends towards the right.
- A **left-skewed distribution** has a "tail" that extends towards the left.
- A **symmetric** distribution has "tails" of approximately equal size.



▲ **FIGURE 2.12** This data set on TV hours viewed per week is skewed to the right. (Source: Minitab Program)

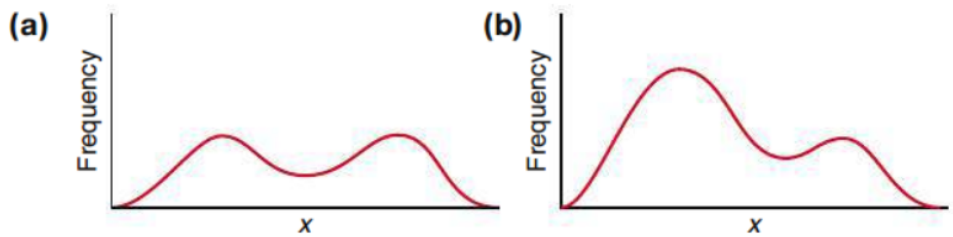


▲ **FIGURE 2.13** This data set on test scores is skewed to the left.

Definition.

- A **unimodal distribution** has data grouped in a single "mound",
- a **bimodal distribution** has data grouped in two "mounds", and
- a **multimodal distribution** has data grouped in more than two "mounds".

▶ **FIGURE 2.14** Idealized bimodal distributions. **(a)** Modes of roughly equal height. **(b)** Modes that differ in height.



Example. In a 5k/10k race where all the runners start at the same time, what do we expect the shape of the distribution of the finishing times will look like?

Definition.
An **outlier** is an extreme value in a distribution of data. Outliers don't fit the pattern of the rest of the data.

Example. Consider the distribution of exam grades. What are possible explanations of any outliers?

Definition.
The most frequently occuring value is called the **mode**.

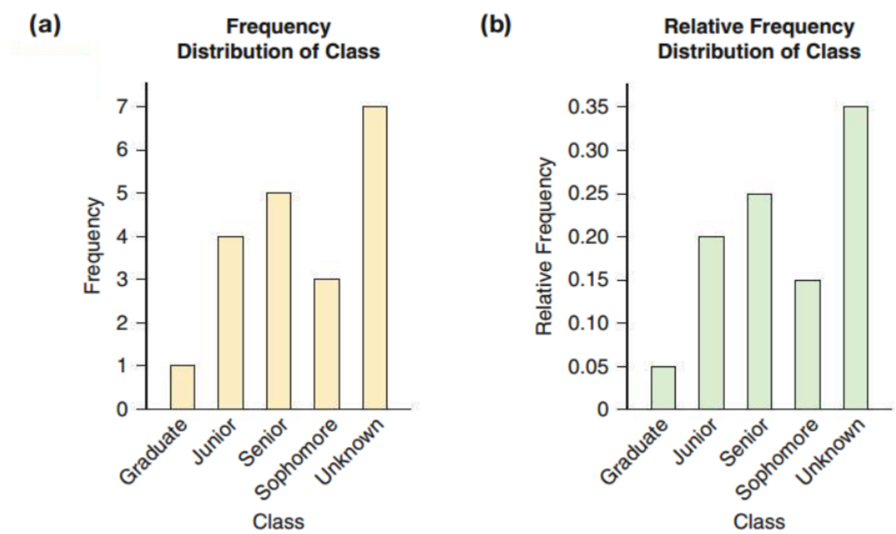
Why might the mode not be a reliable measure of center for numerical data?

2.3: Visualizing Variation in Categorical Variables

Definition.
A **bar chart** (also bar graph or bar plot) shows a bar for each observed category where the height of the bar is proportional to the frequency of that category.

Example. A summer introductory statistics course at UCLA has the following distribution of students across different years:

Class	Frequency
Unknown	7
Freshman	0
Sophomore	3
Junior	4
Senior	5
Graduate	1
Total	20



Bar Charts vs. Histograms:

- Histograms are for numerical data
- Bar charts are for categorical data

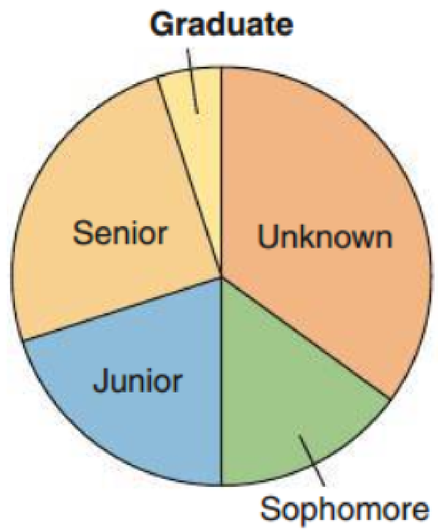
	Histogram	Bar Chart
Bars:	Should touch	May or may not touch
Bar width:	Corresponds to bin width	Can be any width (consistent)
Horizontal labels:	Numerical order	No inherent order

- A **Pareto chart** is a bar graph with bars arranged from tallest to shortest.

Definition.
A **pie chart** is a circle divided up into pieces where each area is proportional to the relative frequency of the category it represents.

Example.

Class	Frequency
Unknown	7
Freshman	0
Sophomore	3
Junior	4
Senior	5
Graduate	1
Total	20



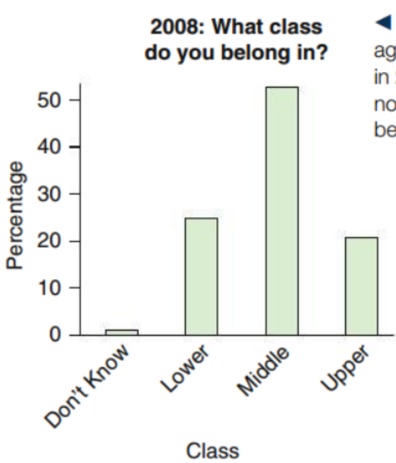
2.4: Summarizing Categorical Distributions

Definition.

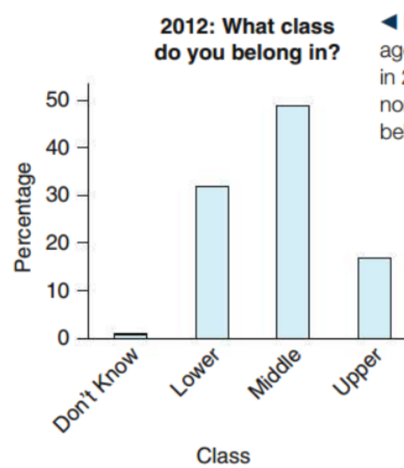
- The category that occurs most often is called the **mode** (similar to the usage with numerical variables).
- A distribution with a lot of *diversity* is said to have a high **variability**.

Note: A categorical variable is considered bimodal *only* if two categories are nearly tied for the mode.

Example. Below are the results of a survey conducted in 2008, and again in 2012. How do the responses compare?

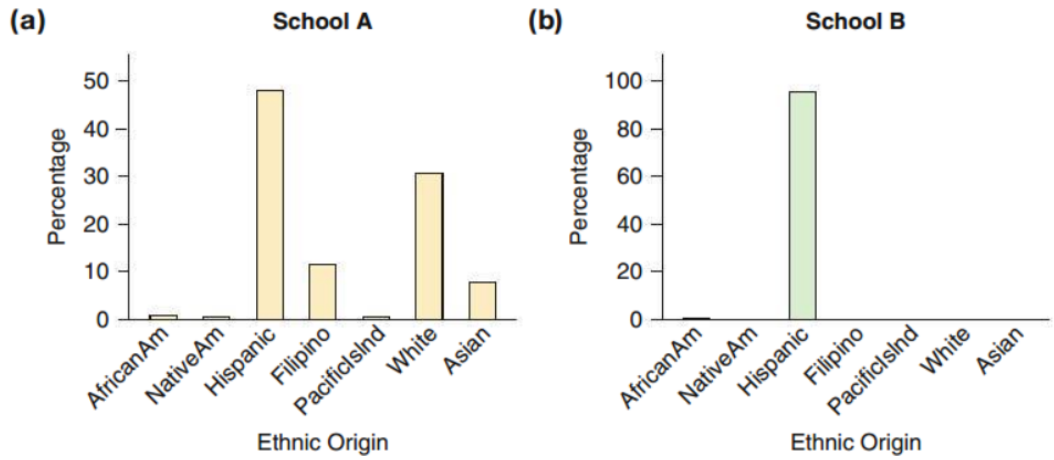


◄ **FIGURE 2.28** Percentages of respondents who, in 2008, reported the economic class they felt they belonged to.



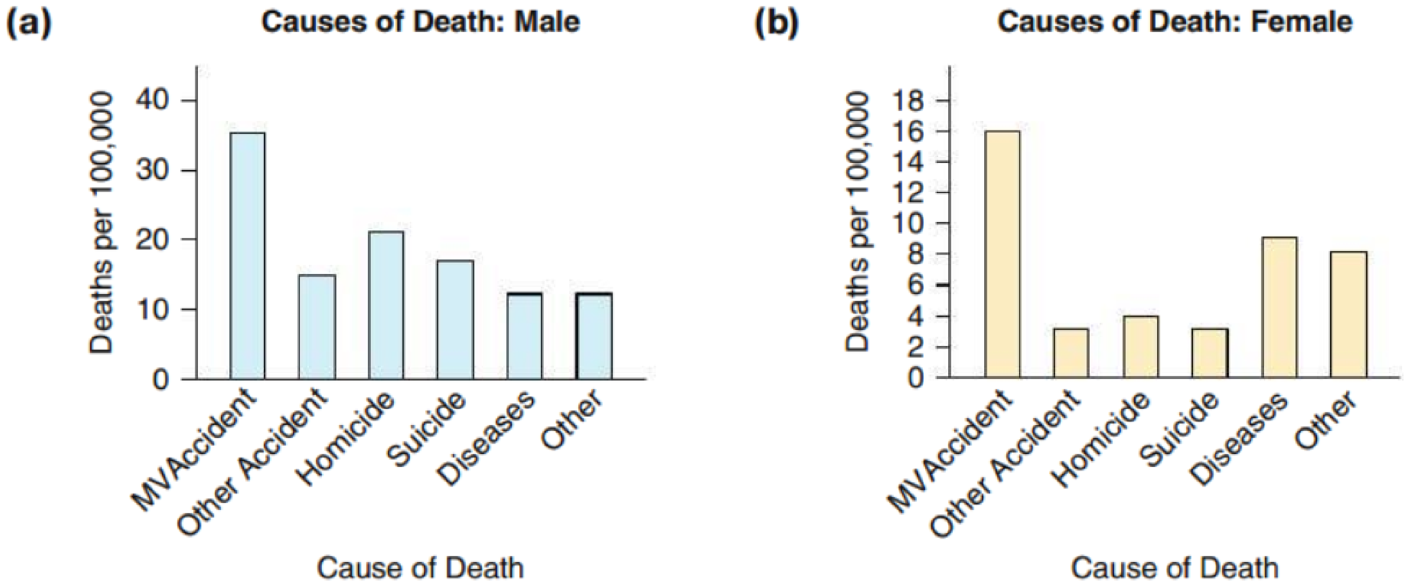
◄ **FIGURE 2.29** Percentages of respondents who, in 2012, reported the economic class they felt they belonged to.

Example. The ethnic composition of two schools in the Los Angeles City School System is presented in the bar charts below. Which school has the greater variability in ethnicity?



◀ **FIGURE 2.30** Percentages of students at two Los Angeles schools who are identified with several ethnic groups. Which school has more ethnic variability?

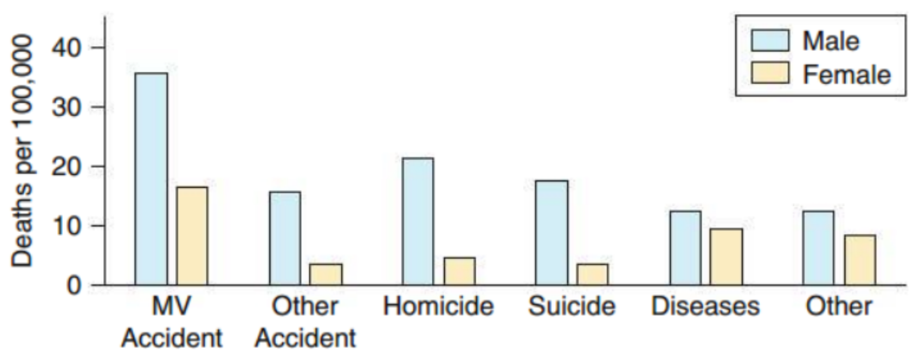
Example. Compare the distributions below. What is the mode for each graph? Which graph demonstrates more variability?



▲ **FIGURE 2.32** The number of deaths per 100,000 males **(a)** and females **(b)** for people 15 to 24 years old in a one-year period.

Example. Compare the combined bar graphs below to the graphs above.

► **FIGURE 2.33** Death rates of males and females, graphed side by side.



2.5: Interpreting Graphs

Appropriate Graphs:
The type of data determines the type of graph you should use

Numerical Data	Categorical Data
Dot plot	Pie chart
Histogram	Bar graph
Stem-and-leaf plot	

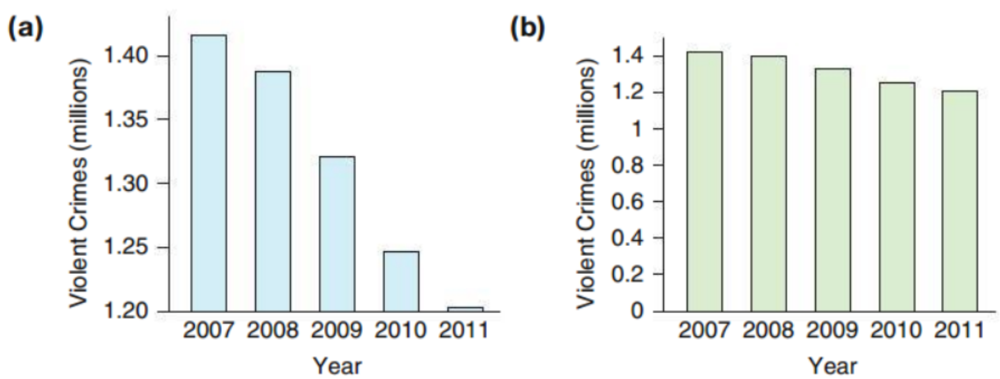
Appropriate Measures:
The type of data determines how the distribution of data should be described

Numerical Data	Categorical Data
Shape	Mode
Center	—
Spread	Variability

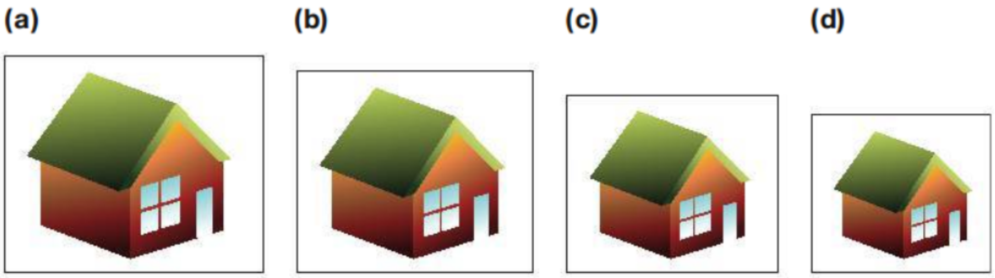
Misleading Graphs: xkcd.com/2023

- Inappropriate scaling (starting at a nonzero value)

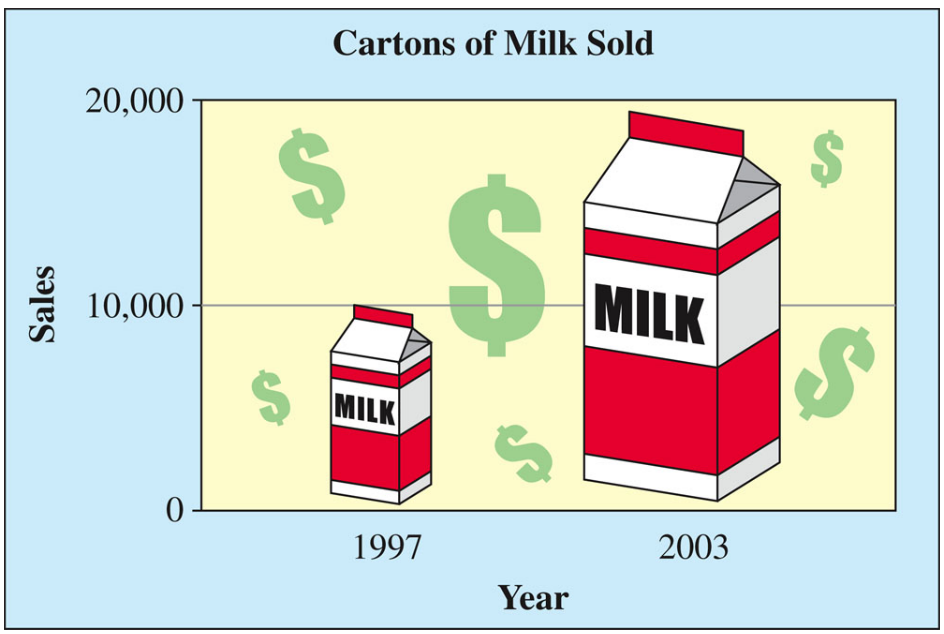
▶ **FIGURE 2.34** **(a)** This bar chart shows a dramatic decline in the number of violent crimes since 2007. The origin for the vertical axis begins at 1.20 million, not at 0. **(b)** This bar chart reports the same data as part (a), but here the vertical axis begins at the origin (0).



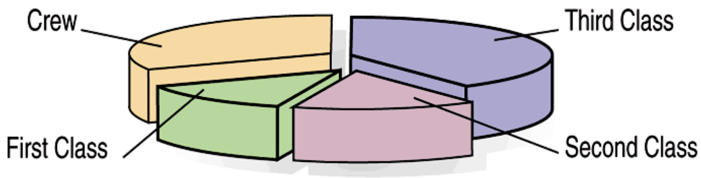
- Icons of different sizes instead of bars of proportionate heights:

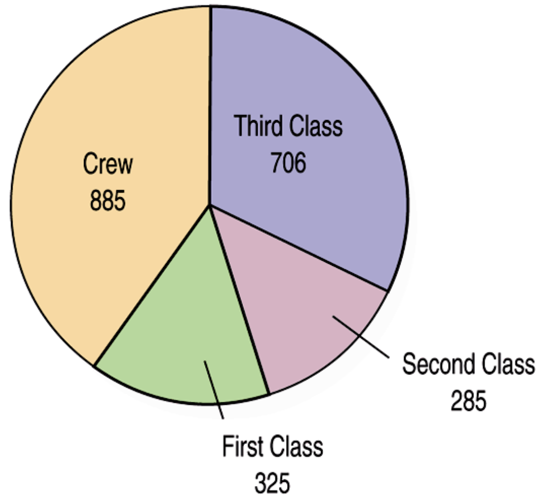


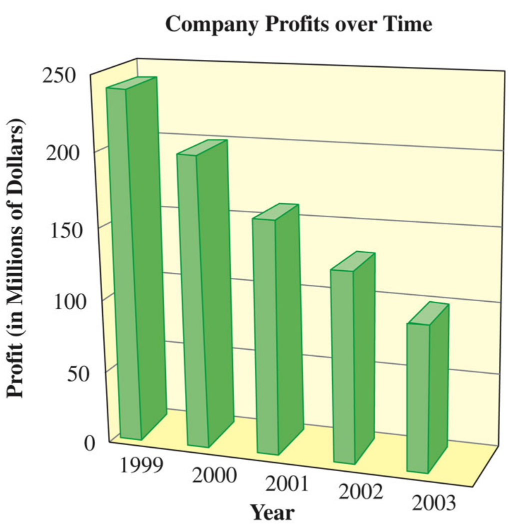
◀ **FIGURE 2.35** Deceptive graphs: Image **(a)** represents 7.1 million homes sold in 2005, image **(b)** represents 6.5 million homes sold in 2006, image **(c)** represents 5.8 million homes sold in 2007, and image **(d)** represents 4.9 million homes sold in 2008. (Source: *L.A. Times*, April 30, 2008)

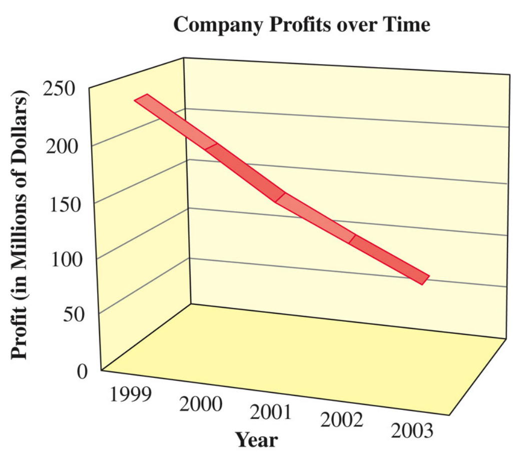


• Avoid the use of 3D graphs:







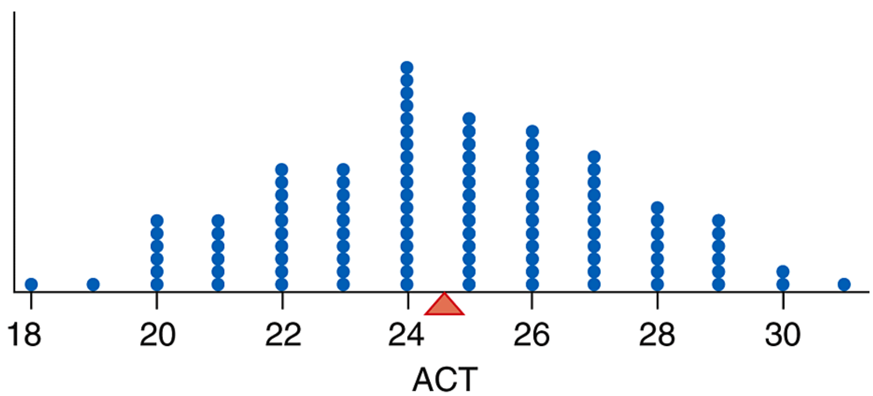


3.1: Summaries for Symmetric Distributions

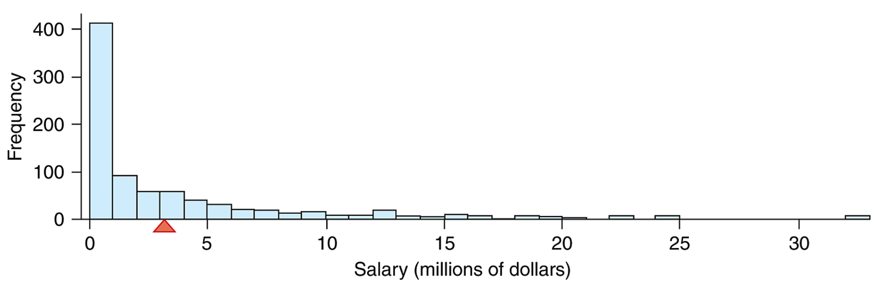
Definition.
Given a collection of data $\{x_1, x_2, \ldots, x_n\}$, the **mean** of the data is the arithmetic mean:

$$\bar{x} = \frac{x_1}{n} + \frac{x_2}{n} + \cdots + \frac{x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example. An instructor at Peoria Junior College in Illinois collected data from two classes, including the students' ACT scores. Below is the distribution of self-reported ACT scores for one statistics class:



Example. The winnings of the top-ranked professional tennis players in the 2018 season are given in the graph below:

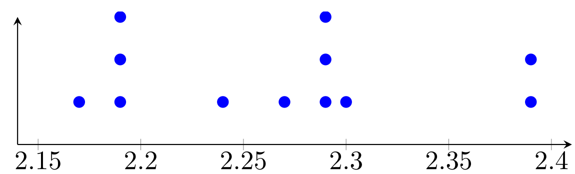


Note:

- When the distribution is roughly symmetric, the mean represents a typical value in the data.
- The mean is *not* a good estimate of a typical value of a skewed distribution.

Example. According to GasBuddy.com (a website that invites people to submit prices at local gas stations), the prices of 1 gallon of regular gas at 12 service stations near the downtown area of Austin, TX, were as follows one winter day in 2018:

\$2.19	\$2.19	\$2.39	\$2.19
\$2.24	\$2.39	\$2.27	\$2.29
\$2.17	\$2.29	\$2.30	\$2.29

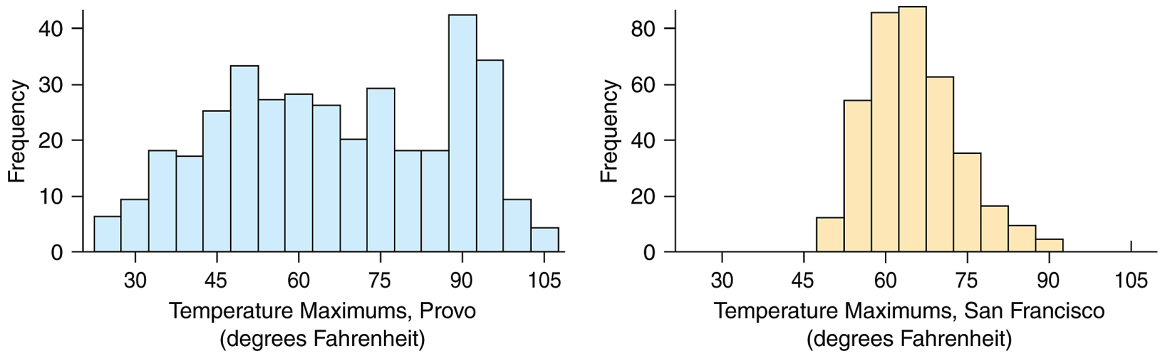


Find the mean price of a gallon of regular gas at these service stations, and interpret the result.

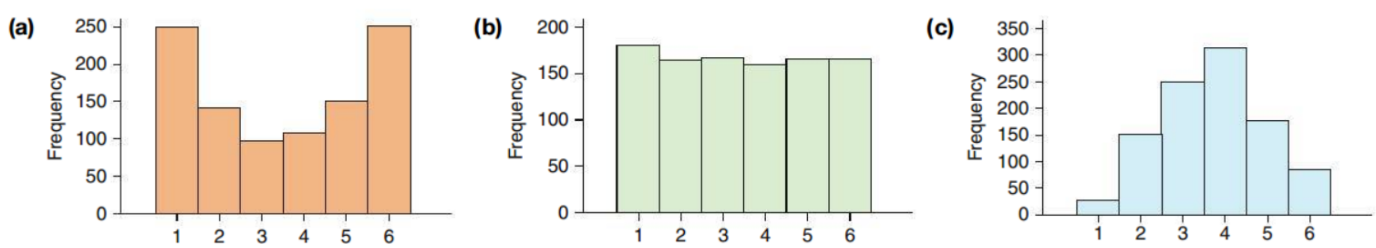
Definition.
The **standard deviation** is a number that measures how far the typical observation is from the mean. For symmetric, unimodal distributions, a majority of the data is within one standard deviation of the mean. The standard deviation is given by

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Example. The histograms below show daily high temperatures in degrees Fahrenheit recorded over one year in Provo, Utah (left), and San Francisco, California (right). Which city do we expect to have a higher standard deviation?



Example. Below are three histograms representing distributions with the same mean. Which distribution has the largest standard deviation? Which has the smallest?



Example. Recall the data set of gas prices from before. Use StatCrunch to compute the standard deviation of this data set and interpret the result: statcrunch.com

\$2.19	\$2.19	\$2.39	\$2.19
\$2.24	\$2.39	\$2.27	\$2.29
\$2.17	\$2.29	\$2.30	\$2.29

1. Click "Open StatCrunch"
2. Enter data into spreadsheet
3. Under the "Stat" menu, select "Summary Stats" then "Columns" or "Rows"

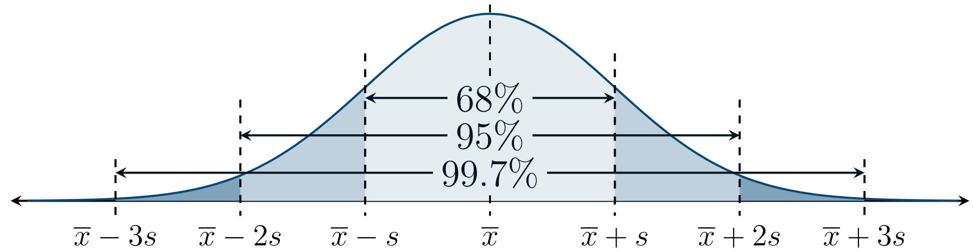
Definition.
The **variance** is the standard deviation squared:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

3.2: What's Unusual? The Empirical Rule and z-Scores

Definition.
The **Empirical Rule** is a guideline for how the standard deviation measures variability. If the distribution is unimodal and symmetric, then

- Approximately 68% of the observations will be within one standard deviation of the mean.
- Approximately 95% of the observations will be within two standard deviations of the mean.
- Nearly all of the observations will be within three standard deviations of the mean.



Example. Recall the data set of gas prices from before.

\$2.19	\$2.19	\$2.39	\$2.19
\$2.24	\$2.39	\$2.27	\$2.29
\$2.17	\$2.29	\$2.30	\$2.29

Recall that the mean gas price was \$2.2666... with a standard deviation of \$0.074. If this is representative of a larger data set, then...

- approximately 68% of the prices would fall between \$2.19 and \$2.34,
- approximately 95% of the prices would fall between \$2.12 and \$2.41, and
- nearly all of the prices would fall between \$2.04 and \$2.49.

Example. The mean daily high temperature in San Francisco is $65°F$ with a standard deviation of $8°F$.

- Find the temperature ranges for 68%, 95%, and 99.7% of the data.

- By the Empirical rule, observations 2 or more standard deviations away from the mean are considered unusual. Is it unusual to have a day when the maximum temperature is colder than $49°F$ in San Francisco?

Example. Suppose that after computing the mean $\bar{x}$ and standard deviation s, we conclude from the empirical rule that approximately 68% of our data lies between 6.5 and 14.78.

- Find the mean $\bar{x}$ and standard deviation s

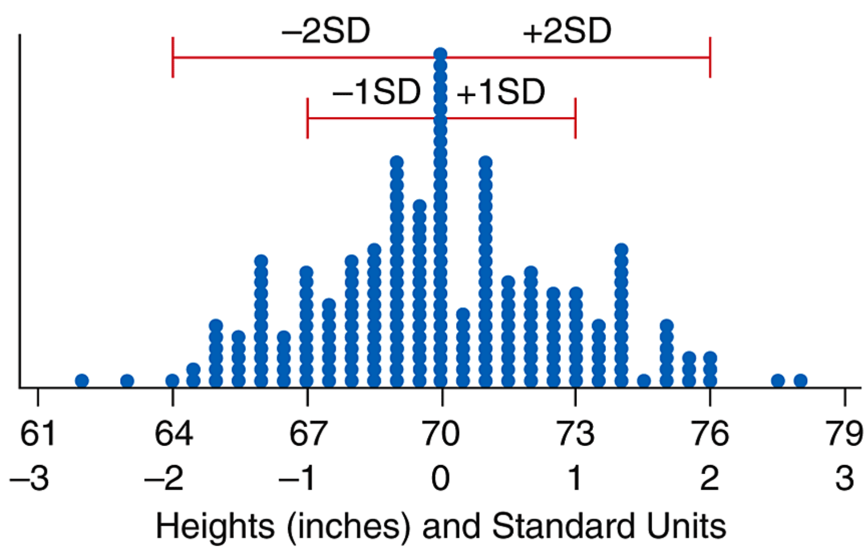
- Use the empirical rule to find the bounds that contain approximately 95% of the data.

> **Definition.**
> A z**-score** measures how many standard deviations an observed data value, x, is from the mean $\overline{x}$:
> $$z = \frac{x - \overline{x}}{s}$$

Example. The dotplot below shows the heights (in inches) of 247 men. The average height is 70 inches, and the standard deviation is 3 inches. How many men have z-scores...

- greater than 2?
- less than -2?

- What is the z-score of a man who is 75 inches tall?



Example. Maria scored 80 out of 100 on her first stats exam in a course and 85 out of 100 on her second stats exam. On the first exam, the mean was 70 and the standard deviation was 10. On the second exam, the mean was 80 and the standard deviation was 5.

On which exam did Maria perform better when compared to the whole class?

3.3: Summaries for Skewed Distributions

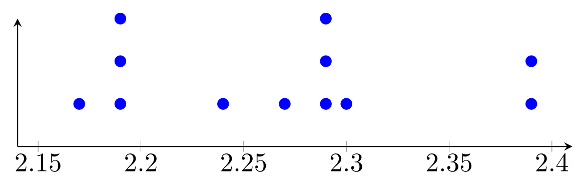
Definition.
The **median** of a sample of data is the middle value when the data is sorted from smallest to largest. If the set contains an

- odd number of observed values, the median is the middle observed value.
- even number of observed values, the median is the average of the two middle observed values.

The median is the preferred measure of center when the data is skewed since about 50% of the observations lie below and above the median.

Example. The prices of 1 gallon of regular gas at 12 service stations near the downtown area of Austin, TX, were as follows one winter day in 2018:

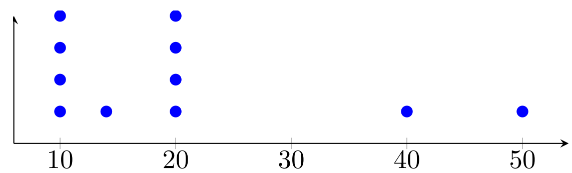
\$2.19	\$2.19	\$2.39	\$2.19
\$2.24	\$2.39	\$2.27	\$2.29
\$2.17	\$2.29	\$2.30	\$2.29



Find the median price for a gallon of gas and interpret the value.

Example. Below are the percentages of fat for some brands of sliced turkey:

$14, 10, 20, 20, 40, 20, 10, 10, 20, 50, 10$



Find the median percentage of fat and interpret the value.

Definition.

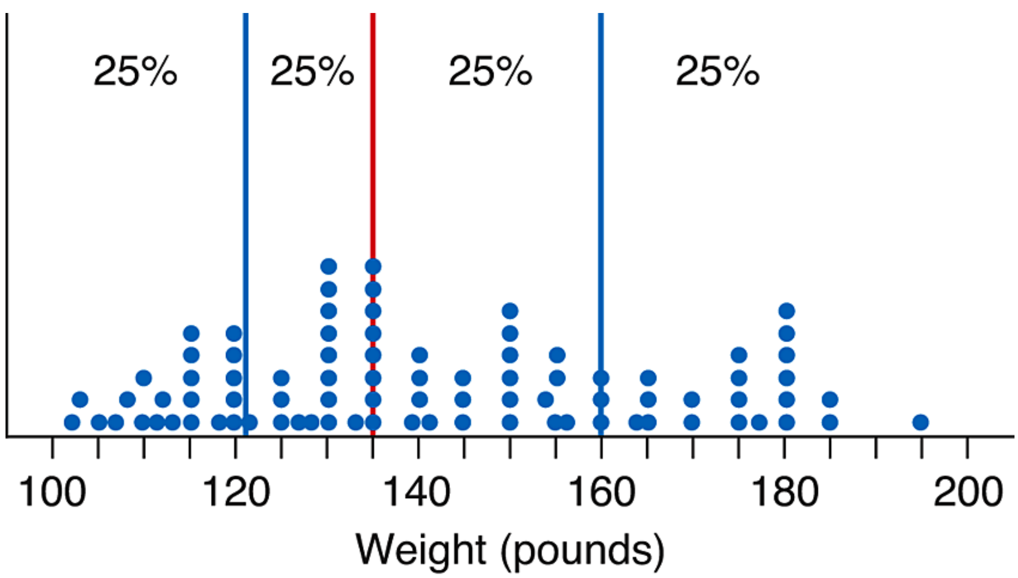
- The **range** is the difference between the maximum and minimum values:

$$\text{Range} = \text{maximum} - \text{minimum}$$

- The **quartiles** divide the data into quarters.
- The **interquartile range (IQR)** indicates approximately how much space the middle 50% of the data occupy.

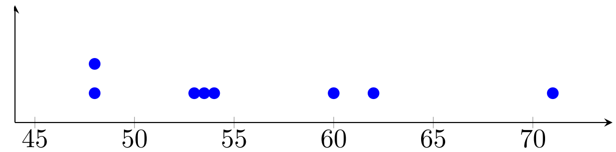
Example. The dotplot below shows the distribution of weights for a class of introductory statistics students.

- Label each line
- Compute the IQR



Example (Computing quartiles by hand)**.** A group of eight children have the following heights (in inches):

$$48.0, 48.0, 53.0, 53.5, 54.0, 60.0, 62.0, 71.0$$



Find the following:

- The median, which is also referred to as Q2

- The first quartile (Q1), which is the median of the lower half of the sorted data

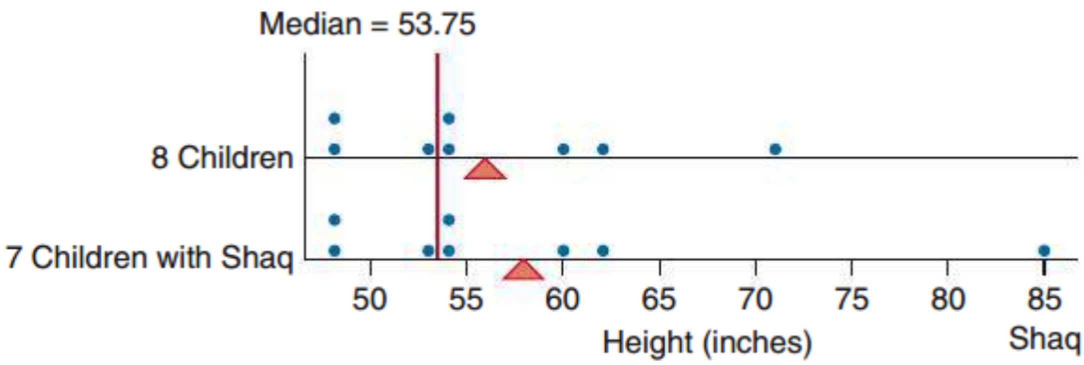
- The third quartile (Q3), which is the median of the upper half of the sorted data

- Compute the IQR

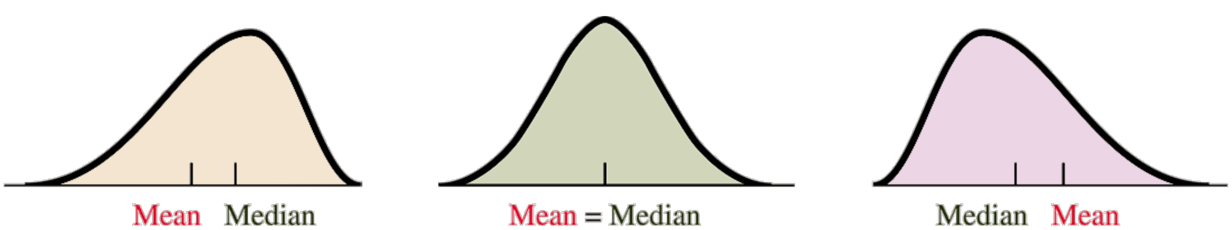
3.4: Comparing Measures of Center

Shape	Measure for center	Measure for spread
Symmetric	Mean	Standard deviation
Skewed	Median	IQR

- Skewed data and outliers affect the mean and standard deviation
- The median is resistant to outliers; it is not affected by the size of an outlier

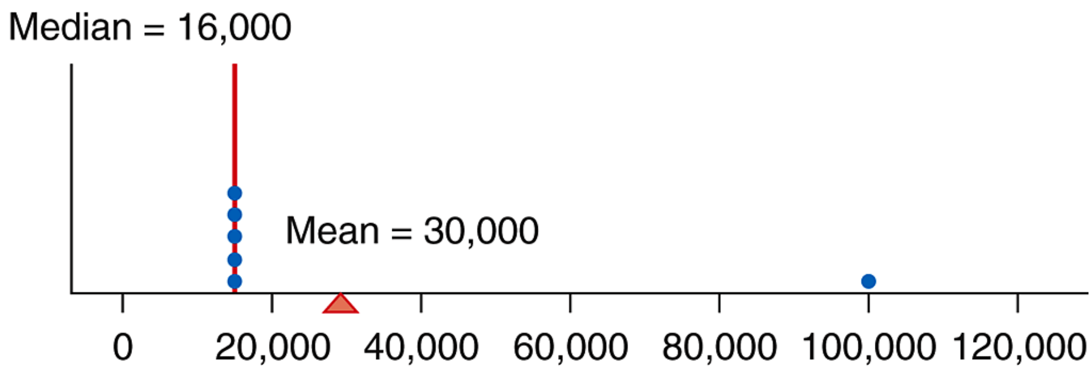


Shape	Mean vs. Median
Skewed left	Mean $<$ Median
Symmetric	Mean $=$ Median
Skewed right	Mean $>$ Median



Example. A (very small) fast-food restaurant has five employees, all of whom work full-time for $7 per hour. Each employee's annual income is about $16,000 per year. The owner, on the other hand, makes $100,000 per year.

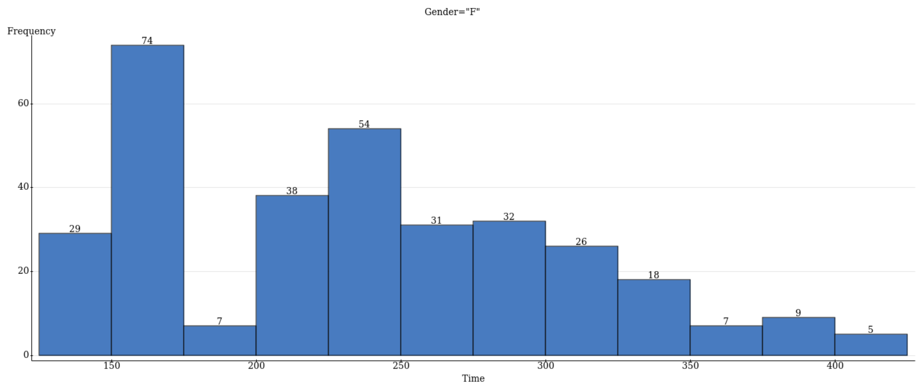
Find both the mean and the median. Which would you use to represent the typical income at this business – the mean or the median? Which value is smaller?



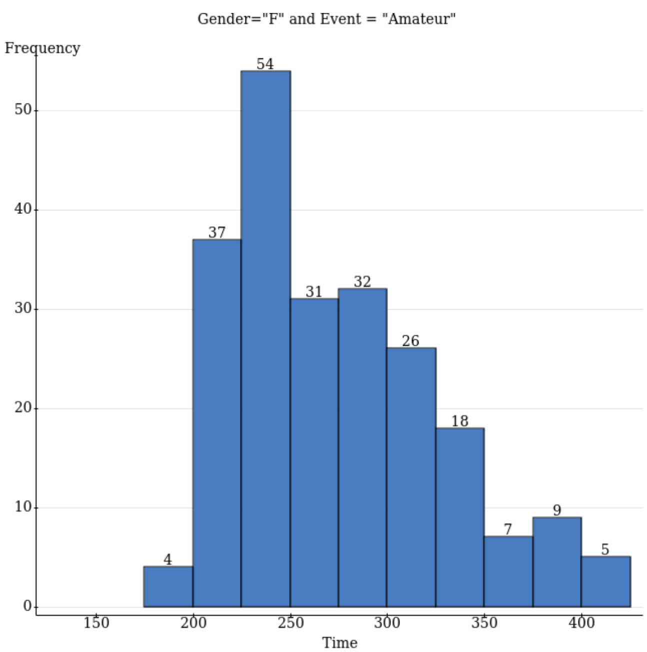
When comparing distributions:

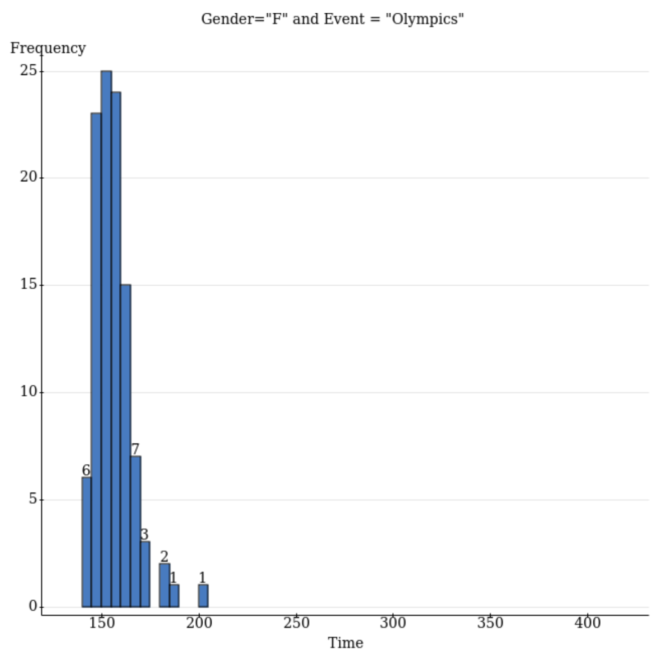
- Always use the same measures of center and spread for both distributions.
- If one of the distributions is skewed, use Median and IQR to compare both!

Example. Below is a histogram of the finishing times of female marathon runners.



If we separate the data into the "Amateur" and "Olympic" events, we see why the data is bimodal. If we compare the distributions, should we use the Mean or the Median? Should we use the standard deviation, or the IQR?

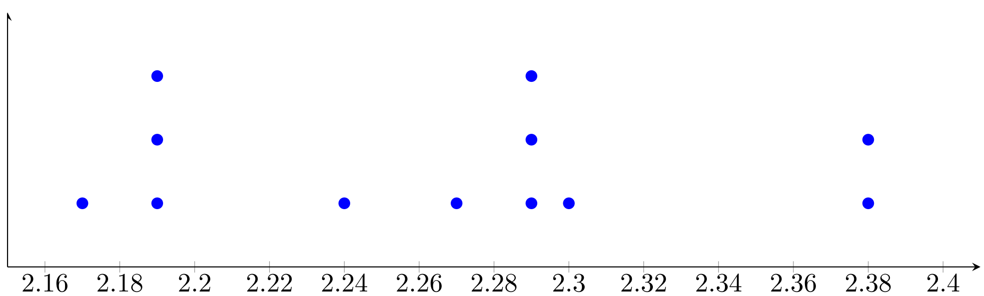


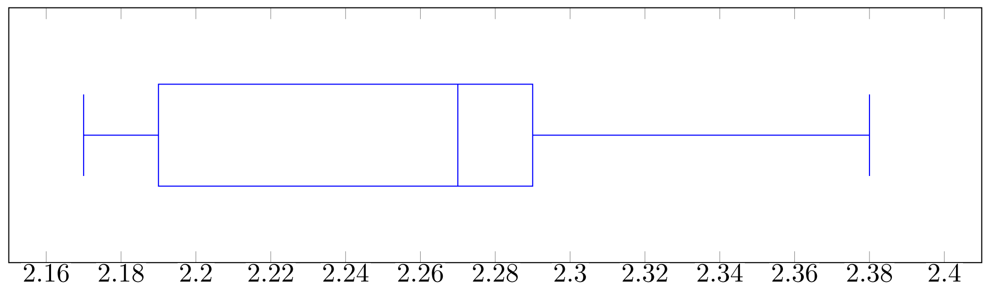


3.5: Using Boxplots for Displaying Summaries

Definition.
A **boxplot** is a graphical tool for visualizing a distribution. Boxplots can be useful for comparing multiple distributions. In a box plot:

- The left edge of the box represents Q_1
- The vertical line inside the box represents the median (Q_2)
- The right edge of the box represents Q_3
- Lines extending past the edges of the box are called whiskers. The whiskers extend to the most extreme values that are not *potential* outliers.



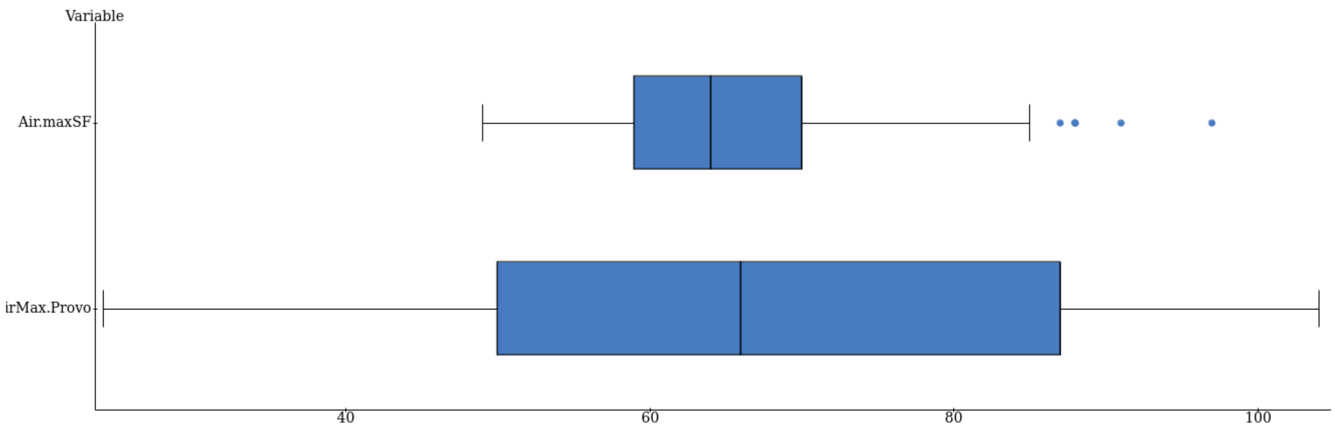


Definition.
Potential outliers are any values that are

- less than $Q_1 - 1.5IQR$
- more than $Q_3 + 1.5IQR$

These values are the left and right limits. They are also known as the *fences*.

Example. Using the "airtemp" dataset in StatCrunch, generate the boxplots for the daily maximum temperature in San Francisco and Provo. Compute the left and right limits. Are they included in the plots?



Definition.
The **five number summary** is

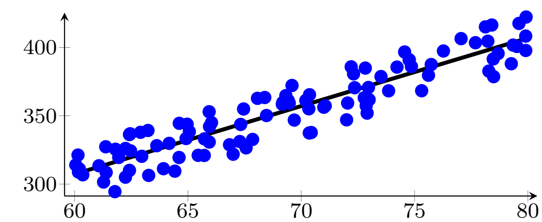
the minimum, Q_1, the median, Q_3, the maximum

4.1: Visualizing Variability with a Scatterplot

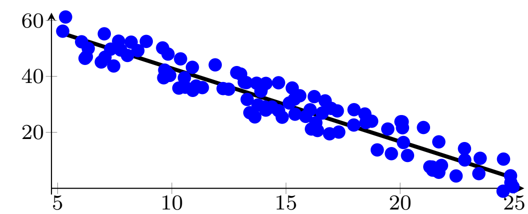
Definition.
A **scatterplot** is used when examining the relationship between two *numerical* variables where each point represents an observation. With scatterplots, we examine the

- **trend**: general tendency of the scatter plot going from left to right
- **strength**: strong associations have little vertical variation
- **shape**: is the scatterplot linear or nonlinear?

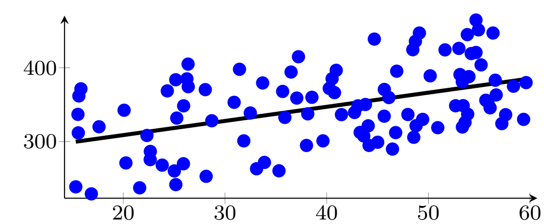
Positive trend



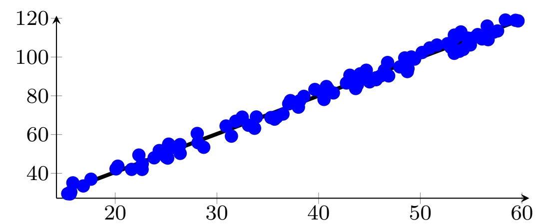
Negative trend



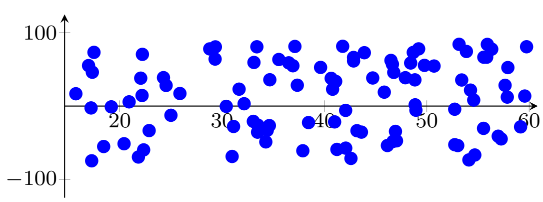
Weak association



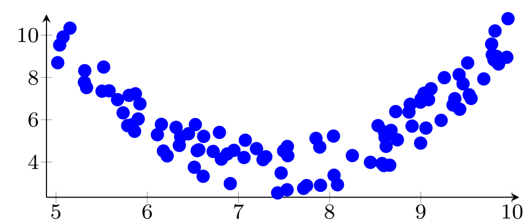
Strong association



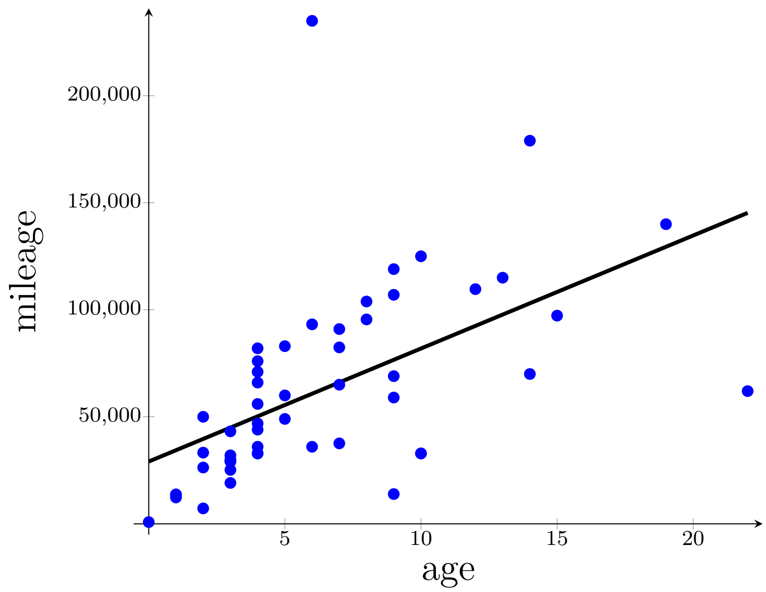
No trend



Quadratic/nonlinear shape



Example. The scatterplot below shows the age and corresponding mileage for a sample of used cars.



What is the association between the variables? Identify the trend, its shape, and how strong the relationship is.

4.2: Measuring Strength of Association with Correlation

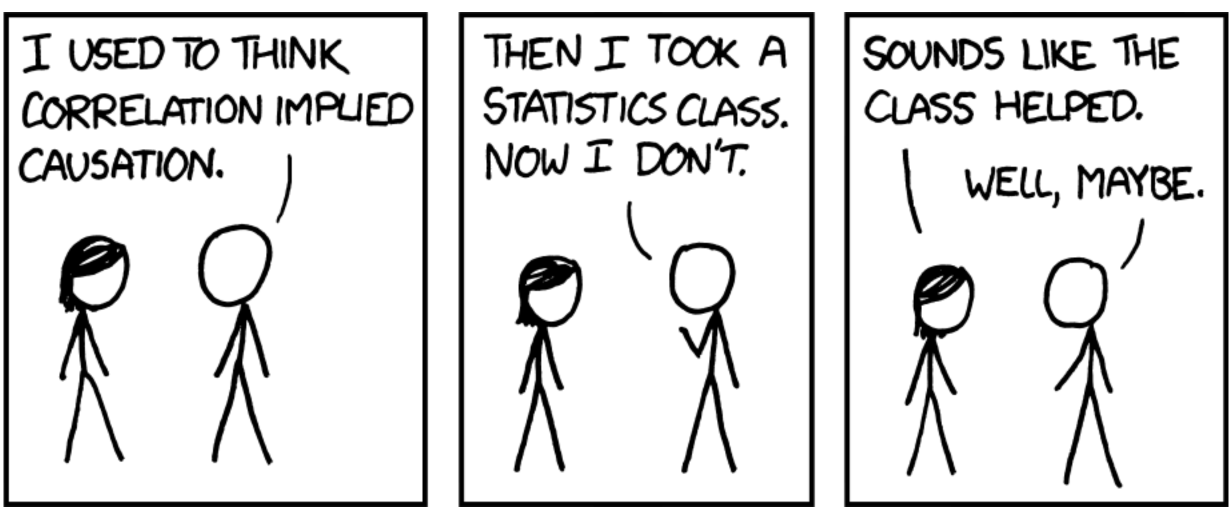
Definition.
The **correlation coefficient** is a number that measures the strength of the linear association between two numerical variables. The correlation coefficient is between -1 and 1:

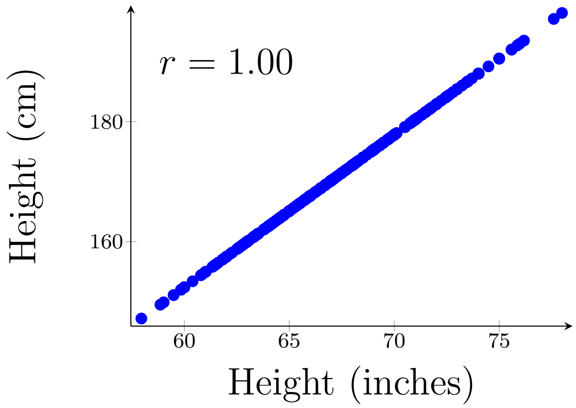
$1 \rightarrow$	Strong association and positive trend
$0 \rightarrow$	Weak or no association
$-1 \rightarrow$	Strong association and negative trend

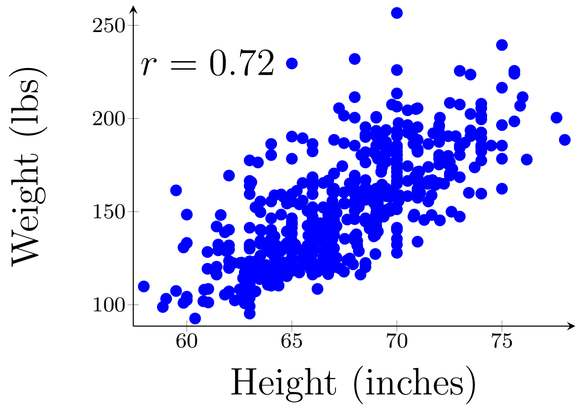
The correlation coefficient only makes sense if the trend is linear and both variables are numerical!!

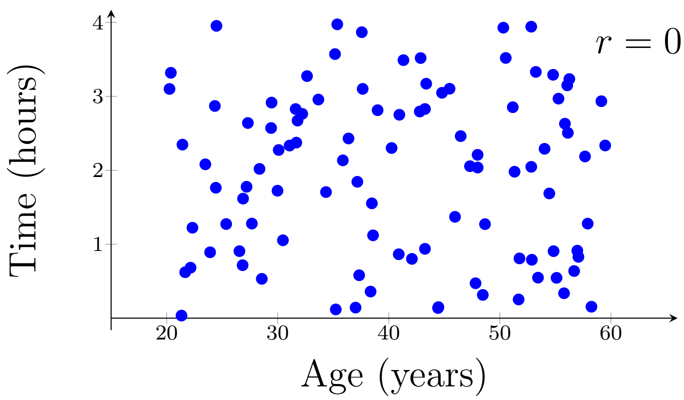
Note: *Correlation does not mean causation!*

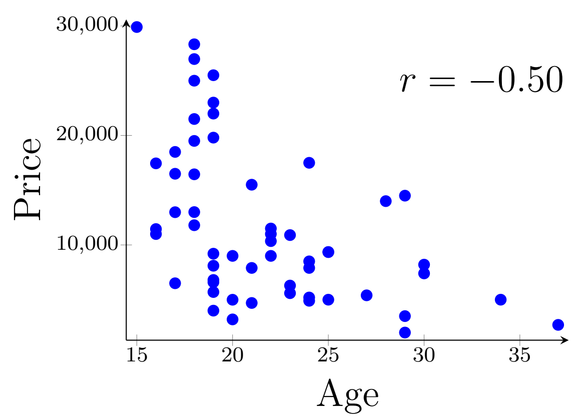
Take a few minutes to look at the graphs at this link: Spurious correlations

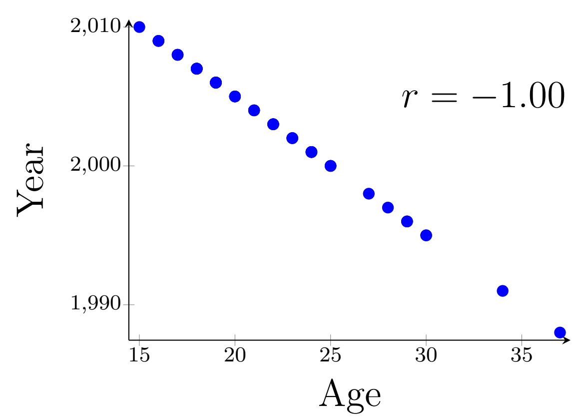



$r = 1.00$
Height (cm)
180
160
60
65
70
75
Height (inches)


$r = 0.72$
Weight (lbs)
250
200
150
100
60
65
70
75
Height (inches)


$r = 0$
Time (hours)
4
3
2
1
20
30
40
50
60
Age (years)


$r = -0.50$
Price
30,000
20,000
10,000
15
20
25
30
35
Age


$r = -1.00$
Year
2,010
2,000
1,990
15
20
25
30
35
Age

Definition.
The formula for the correlation coefficient between two variables x and y is

$$r = \frac{\sum z_x z_y}{n-1}$$

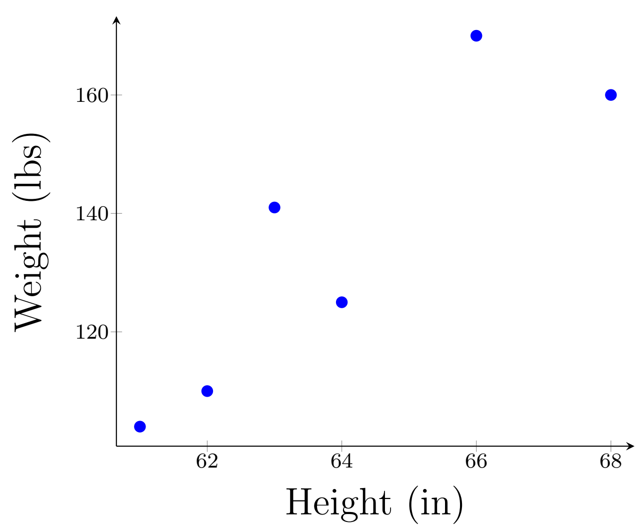
where z_x and z_y are the z scores for each entry in the x and y lists.

$$z_x = \frac{x - \overline{x}}{s_x} \qquad z_y = \frac{y - \overline{y}}{s_y}$$

Example. Below are the heights and weights of six women:

Heights	61	62	63	64	66	68
Weights	104	110	141	125	170	160

Compute the correlation coefficient by hand. Then, graph the scatterplot and compute the correlation coefficient using StatCrunch.



Understanding the Correlation Coefficient:

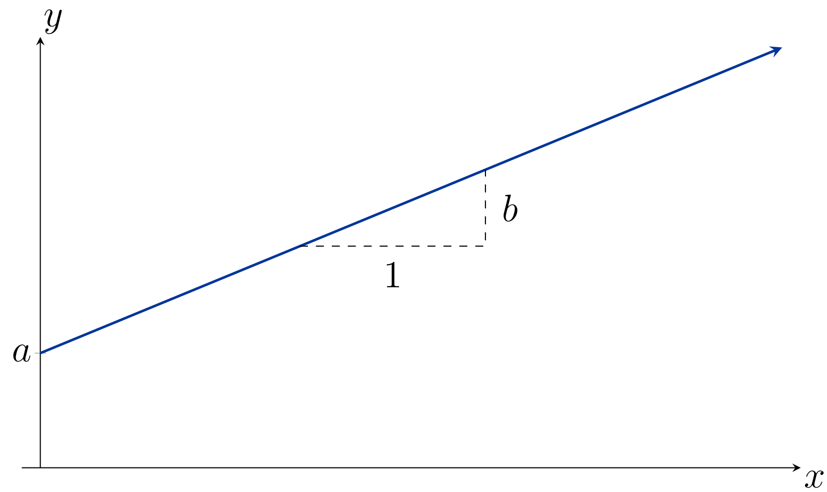
- Changing the order of the variables does not change r
- Adding a constant or multiplying by a positive constant does not affect r
- The correlation coefficient is unitless
- None of this makes any sense if the relationship between the variables is not linear!

4.3: Modeling Linear Trends

Definition.
The **regression line** is a model used for making predictions about *future* observed values. The equation of the regression line is

$$y = a + bx$$

where a is the y**-intercept** and b is the **slope**.



The input variable x is known as the

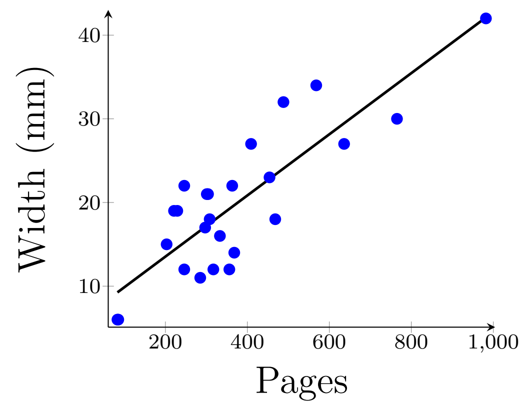
- Independent variable
- Predictor variable
- Explanatory variable

The output variable y is known as the

- Dependent variable
- Predicted variable
- Response variable

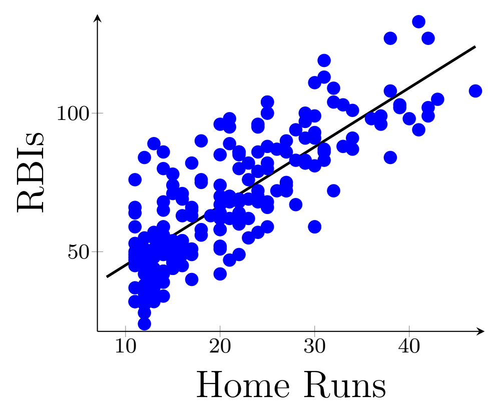
Example. Below is a scatterplot comparing number of pages a book has against the width of the book. Interpret the intercept and the slope of the regression line.

Predicted Width=6.22+0.0366 Pages



Example. Below is a scatterplot comparing the number of home runs and RBIs in the 2016 season. Interpret the intercept and slope of the regression line.

Predicted RBI=23.84+2.13 HR



Definition.

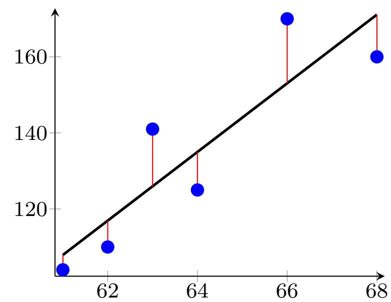
Now we define the formula of the regression line:

$$y = a + bx$$

Where

$$b = r\frac{s_y}{s_x} \quad \text{and} \quad a = \overline{y} - b\overline{x}.$$

These formulae minimize the residual error: Try this!



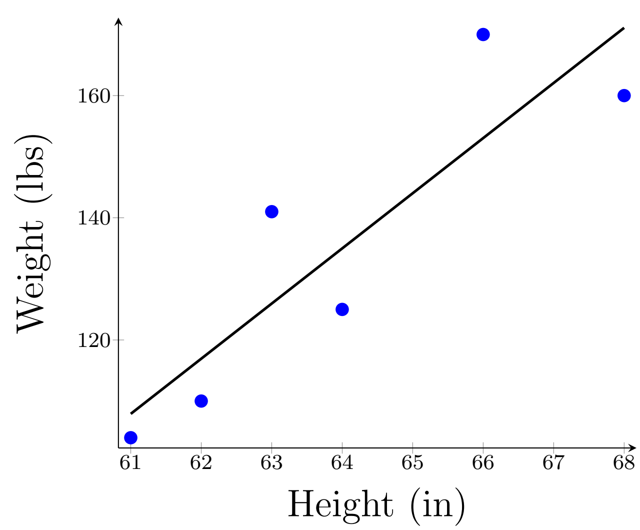
Example. Below are the heights and weights of six women:

Heights	61	62	63	64	66	68
Weights	104	110	141	125	170	160

From this we get

$$\overline{y} = 135 \qquad s_y = 26.728$$
$$\overline{x} = 64 \qquad s_x = 2.608$$
$$r = 0.881$$

Find the equation of the regression line.



Example. Open the `popdensity_and_crime` dataset in StatCrunch, use the "Simple Linear" tool under the `Stat>Regression` menu to find the regression line for the following columns. Interpret the slope and intercept where appropriate.

the `pop1990` and `pop2000` columns,

the `pop2000` and `totcrimerate` columns, and

the `pop2000` and `Rank Pop` columns.

4.4: Evaluating the Linear Model

Guidelines:

- Don't fit linear models to nonlinear associations!
- Correlation is not causation
- Beware of outliers (a.k.a. **influential points**)
- Don't extrapolate (make predictions beyond the range of the data)

Definition.
The **coefficient of determination** is the correlation coefficient squared:

$$r^2$$

This is sometimes also called r**-squared**.

6.1: Probability Distributions Are Models of Random Experiments

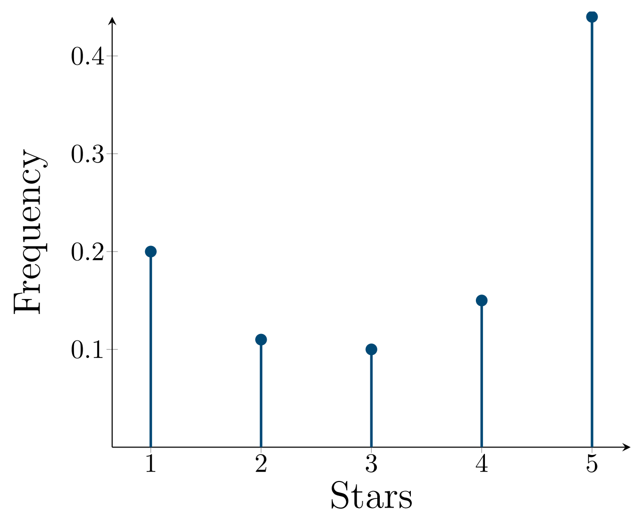
Definition.
The **probability distribution** describes

- all possible outcomes of a random experiment, and
- the probability of each outcome.

This is sometimes also referred to as the **probability distribution function (pdf)**.

Example. Suppose we are reading Amazon reviews of a particular product. In total, the product has $3,901$ reviews, distributed as shown below.

Stars	Frequency
5	0.44
4	0.15
3	0.10
2	0.11
1	0.20



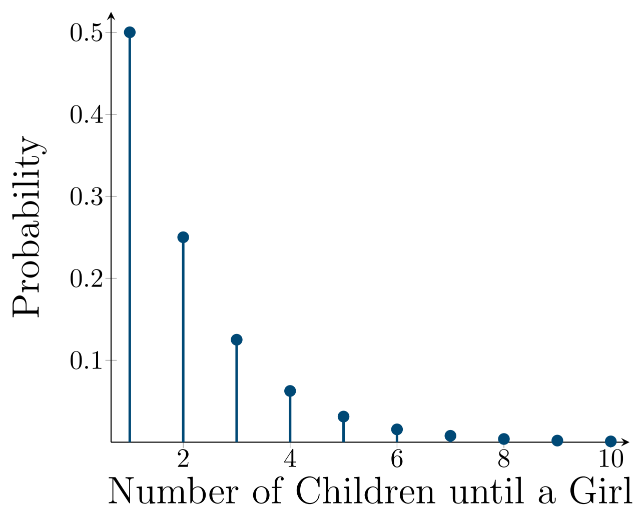
If we pick a reviewer at random, what is the probability that they give a 5 star review? What about a 1 star review?

What is the sum of the probabilities?

Note: Valid probability distributions:

- Have probabilities between 0 and 1,
- The sum of the probabilities is *exactly* 1.

Example. Suppose a couple decides they will keep having children until they have a girl. Assuming that the likelihood of having a boy or girl is equally likely, the probability of having x children can be given by $(1/2)^x$, and is represented by the graph below.



What is the maximum number of children possible?

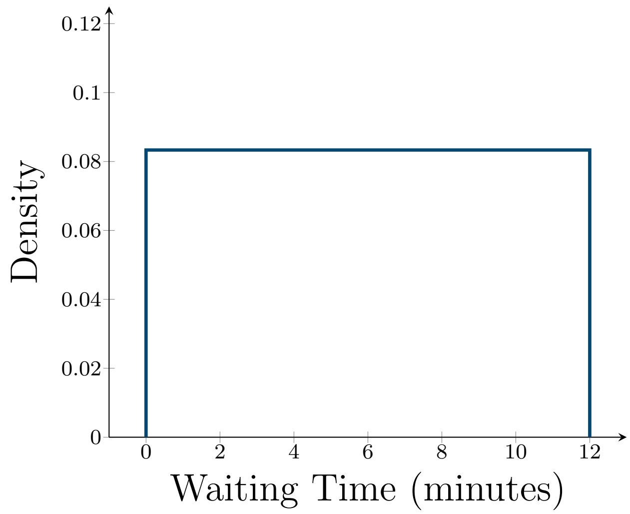
Do the probabilities sum to 1?

Finding the probabilities for continuous outcomes:

- is represented as area under a curve,
- is in the context of a range of values, and
- the probability of hitting an exact value is 0

Example. Suppose a coffee shop has done extensive research and knows each customer is helped in under 5 minutes. The shaded area of the graph represents the probability that a customer will wait less than 2 minutes.

Example. Suppose a bus arrives at the bus stop every 12 minutes. If you arrive at the bus stop at a randomly chosen time, then the probability distribution for the number of minutes you must wait is shown in the graph below:



Find the probability that you will have to wait less than 5 minutes.

Find the probability that you will have to wait between 4 and 10 minutes.

What is the probability that you will have to wait *exactly* 12 minutes?

6.2: The Normal Model

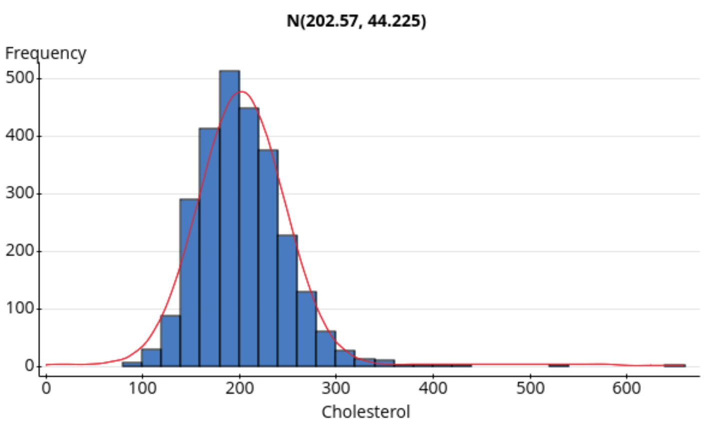
Definition.
The **Normal Distribution** is a symmetric, unimodal model that provides a very close fit for many numerical variables:

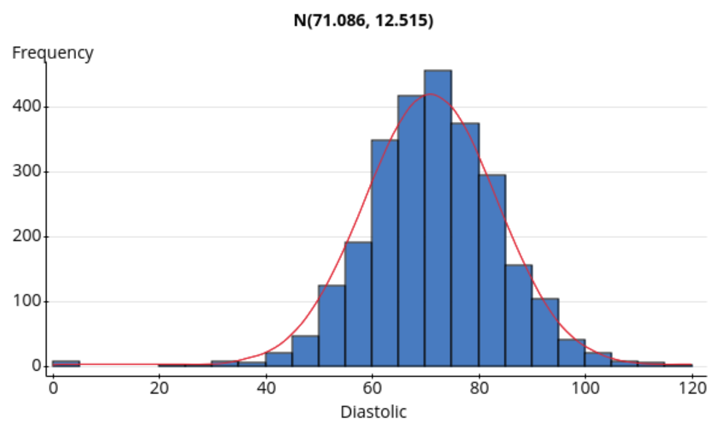
$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

We use $N(\mu, \sigma)$ to denote the Normal Distribution with mean μ and standard deviation σ.
Note:

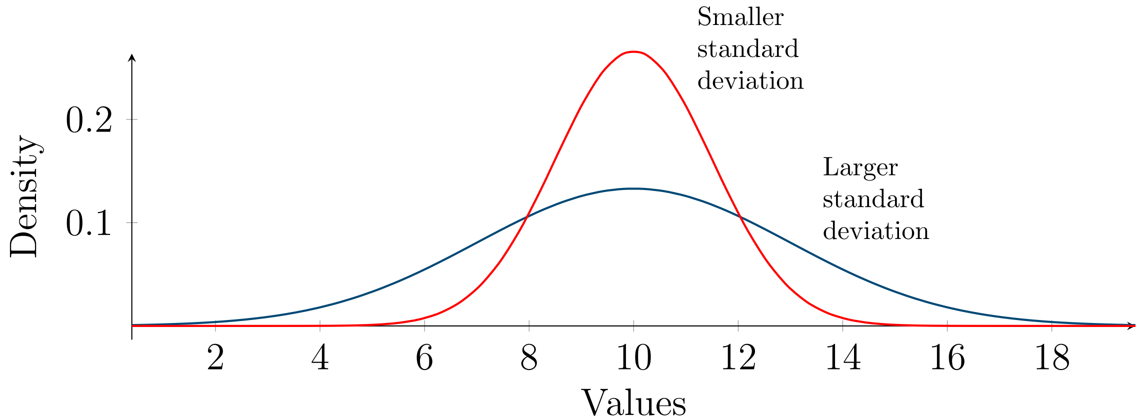
- μ and σ are used in the context of a probability distribution, whereas $\overline{x}$ and s are used for data.
- Other sources denote the Normal Distribution with mean μ and *variance* σ^2 as $N(\mu, \sigma^2)$ or $\mathcal{N}(\mu, \sigma^2)$.

Example. Below are some histograms from a dataset that show the measured cholesterol and diastolic blood pressure from 2,793 people. These histograms have the Normal Distribution with the corresponding mean μ and standard deviation σ overlayed:

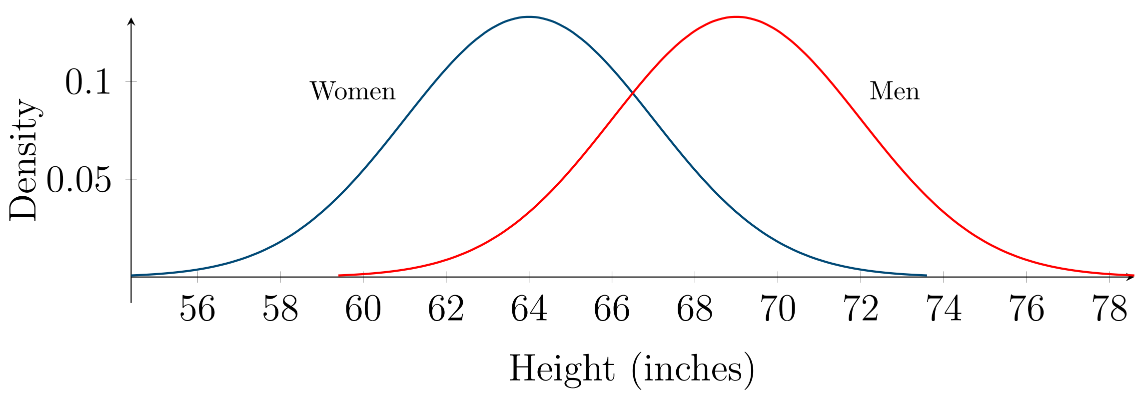




Example. Below is the graph of two Normal Distributions with equal means, but different standard deviations.



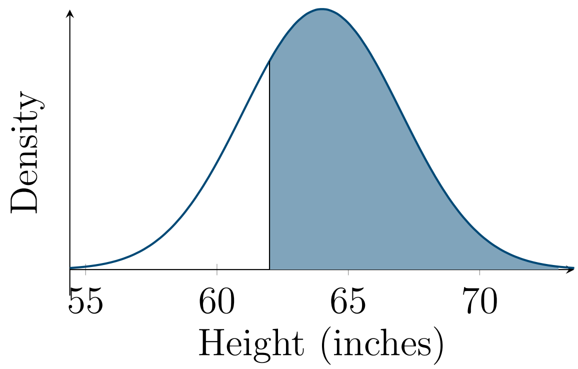
Example. Below is the graph of two Normal Distributions with equal standard deviations, but different means.



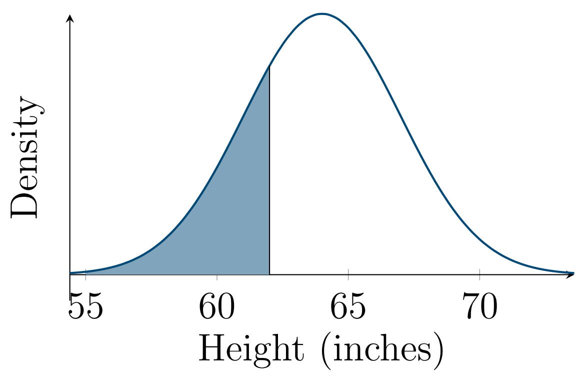
What is the area under each of the curves above?

Example. Suppose that the Normal model $N(64, 3)$ gives a good approximation to the distribution of adult women's height in the United States. If a women is chosen at random, what is the probability that

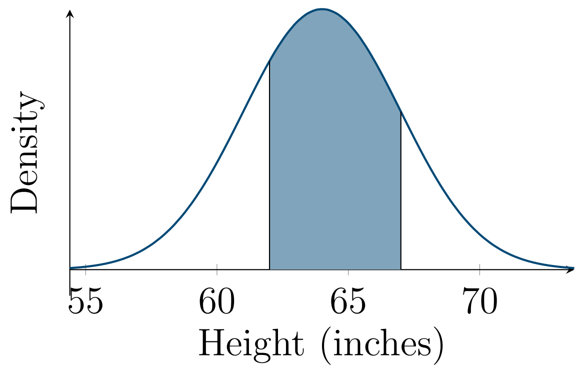
she is taller than 62"?



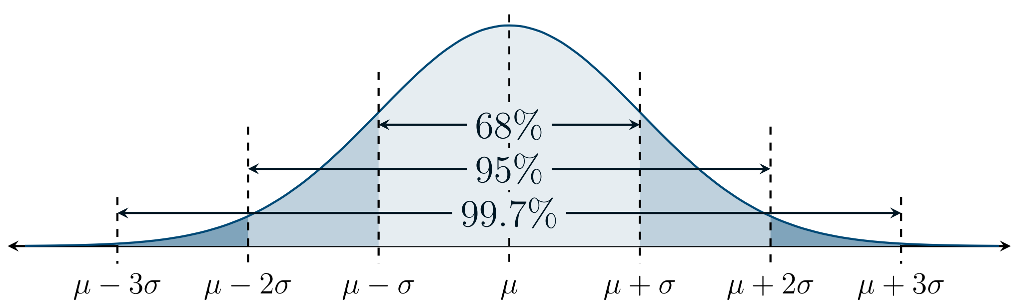
she is shorter than 62"?



her height is between 62” and 67”?



Example. “Verify” the emperical rule by using technology to find the probability that an observation lies within 1, 2, and 3 standard deviations.



Definition.
The **Standard Normal Distribution** is a $N(0,1)$:

$$\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

We use the Standard Normal Distribution in conjunction with z-scores to compute probabilities:

$$z = \frac{X-\mu}{\sigma}$$

Example. Suppose the length of a newborn seal pup follows a Normal Distribution with a mean length of 29.5, and standard deviation 1.2. Solve the following by finding the z-score and then using a z-score table to compute the probability that a seal pup's length is

shorter than 28”, longer than 31”, and

is between 28” and 31”.

Example. Assume that women's heights follow a Normal distribution with mean 64" and standard deviation 3". Find the 25th and 75th percentile using

technology and

by hand.

7.1: Learning about the World through Surveys

Definition.
A **population** is a group of objects or people we wish to study.

- A **parameter** is a numerical value describing some aspect of the population (e.g. means and proportions)
- A **census** is a survey of *every member* of a population

A **sample** is a subset of the population of interest.

- A **statistic** is a numerical value describing some aspect of the sample
- Statistics are sometimes called **estimators**
- A **statistical inference** is the science of drawing conclusions about a population based on observing only a small subset of that population.

Example. In February 2014, the Pew Research Center surveyed 1428 cell phone users in the United States who were married or in a committed partnership. The survey found that 25% of cell phone owners felt that their spouse or partner was distracted by their cell phone when they were together.

Identify the population and the sample

Identify the parameter and the statistic

Sample statistics and population parameters are represented using different symbols (English for statistics, Greek for population):

	Sample	Population
	Statistics	Parameters
mean	$\overline{x}$	μ
standard deviation	s	σ
proportion	$\hat{p}$	p

Example. The City of Los Angeles provides an open data set of response times for emergency vehicles. Each row of the data set represents an emergency vehicle that has been sent to a particular emergency. A random sample of 1000 of these rows shows that the mean response time was 8.25 minutes. In addition, the proportion of vehicles that were ambulances was 0.328.

Using correct notation, identify the data given above.

What can we conclude about the overall population?

Definition.
A method is **biased** if it tends to produce the wrong value.

- **Sampling bias** results from a sample that is not representative of the population.
- **Measurement bias** results from questions that do not produce a true answer.

Example. For the following scenarios, identify any bias:

Online reviews (Amazon, Yelp, etc.)

Asking if people support a 'fat tax' on non-diet sugary soft drinks

Gallup poll calling landline phones

Using a poorly written question (e.g. double negative)

Definition.
A **simple random sample (SRS)** is where subjects from a population are drawn *at random* and *without replacement.* With an SRS, each member of the population has an equally likely chance of being selected.
Nonresponse bias results from people refusing to respond to the survey.

Example. Perform an SRS of 3 people from the list below:

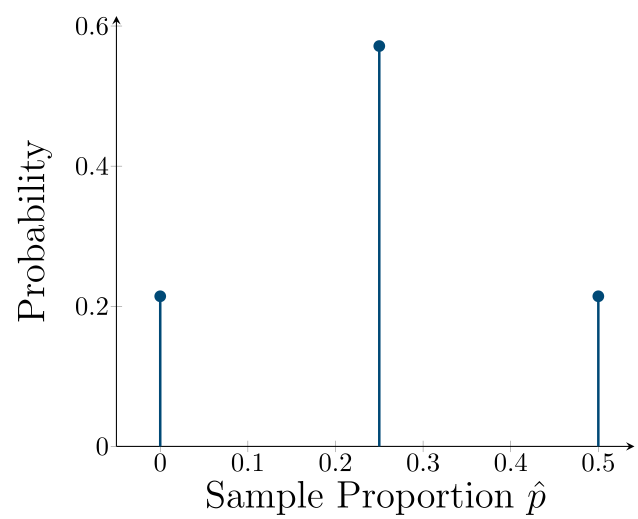
1	Alberto
2	Justin
3	Michael
4	Audrey
5	Brandy
6	Nicole

7.2: Measuring the Quality of a Survey

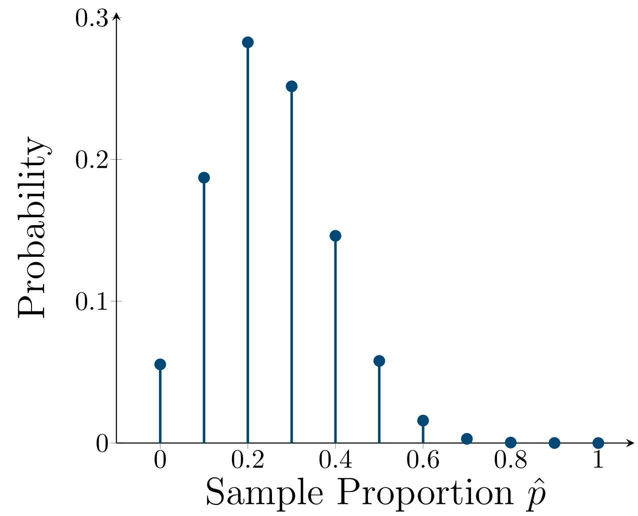
The true population proportion can be estimated by the sample proportion. How accurate can we expect our estimate to be?

- The *accuracy* of an estimation method is measured in terms of *bias*
- The *precision* of an estimation method is measured in terms of *standard error*

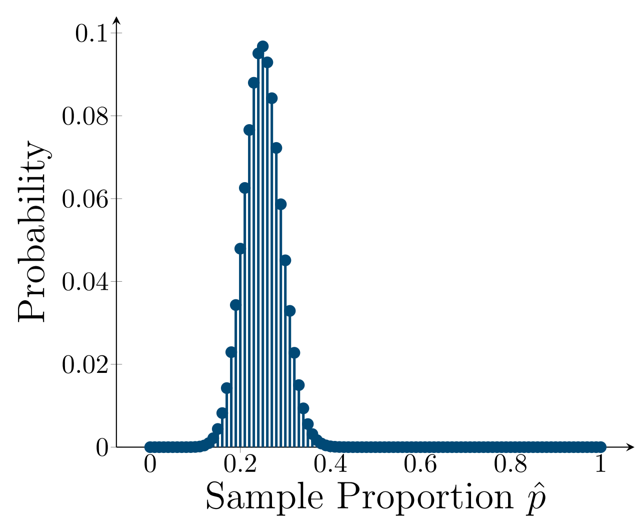
Example. Consider a group of 8 people, where 2 identify as female, and 6 identify as male. What is the true population proportion of females? When using a sample size of $n = 4$, what are possible sample proportions?



Example. Now, consider a group of 1000 people where 25% identify as female ($p = 0.25$). When using a sample size of $n = 10$, what are possible sample proportions?



Finally, consider a group of 1000 people where 25% identify as female ($p = 0.25$). When using a sample size of $n = 100$, what are possible sample proportions?



Definition.

- The **sampling distribution** is the probability distribution of $\hat{p}$.
- The **standard error** for $\hat{p}$ is given by

  $$SE = \sqrt{\frac{p(1-p)}{n}}$$

  provided that

  - The sample is randomly selected from the population of interest.
  - If sampling without replacement, the population needs to be much larger than the sample size (e.g. at least 10 times bigger)

  Since the true population proportion is typically unknown, we can estimate the standard error:

  $$SE_{est} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Note: Larger sample sizes have smaller standard error!

7.3: The Central Limit Theorem for Sample Proportions

Definition. (Central Limit Theorem (CLT))

When estimating a population proportion, p, if

1. *Random and Independent*: The sample is collected randomly from the population, and observations are independent of each other.
2. *Large Sample*: The sample size, n, is large enough that the sample can have at least 10 successes or failures.
$$\begin{aligned} n\hat{p} &\geq 10 \\ n(1-\hat{p}) &\geq 10 \end{aligned}$$
3. *Big population*: If the sample is collected without replacement (e.g. SRS), then the population size must be at least 10 times bigger than the sample size.
$$N \geq 10n$$

then the sampling distribution for $\hat{p}$ is approximately Normal, with mean p and standard deviation
$$SE = \sqrt{\frac{p(1-p)}{n}}.$$
This distribution is denoted as
$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

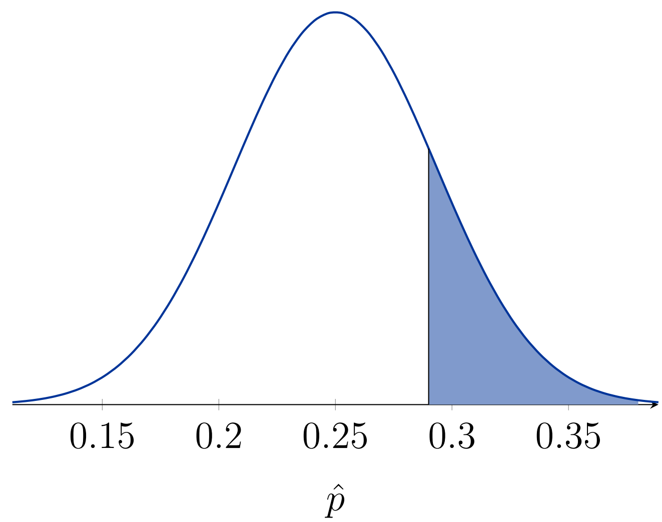
Example. Consider the groups from the previous section where $p = 0.25$ of the group identified as female. Suppose that $\hat{p} = 0.25$. If N represents the population size, and n the sample size, identify if the CLT can be applied.

$N = 8, n = 4$

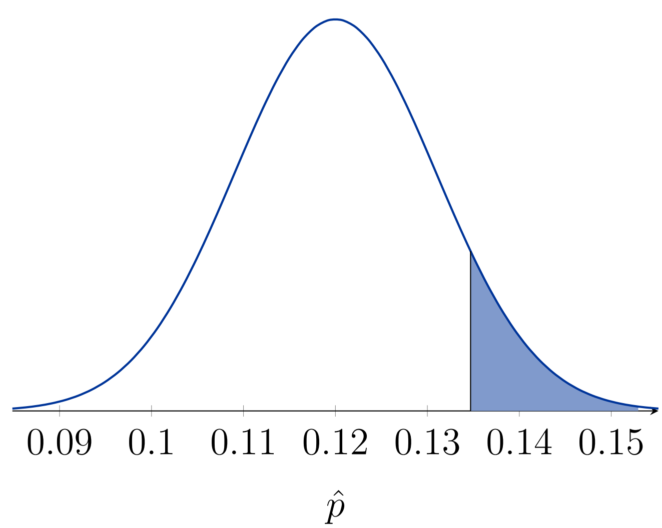
$N = 1000, n = 10$

$N = 1000, n = 100$

Example. Consider the group of 1000 people where $p = 0.25$ identified as female. In a sample of $n = 100$ people, what is the probability that $\hat{p}$ is at least 29%?



Example. Samuel Morse claimed that the true proportion of E's used in the English language is 0.12. Suppose we take a sample of 876 letters, and find a sample proportion of 0.1347. If we took another sample, what is the probability that the new sample proportion would be greater than 0.1347?



7.4: Estimating the Population Proportion with Confidence Intervals

Definition.
Suppose that we wish to estimate a population proportion p based on a sample proportion $\hat{p}$.

- A **confidence interval** is an interval about the point estimate $\hat{p}$ that we can be confident contains the true population proportion p:

$$\hat{p} \pm m$$

- The **margin of error (ME)** is half the width of the confidence interval. When estimating a population proportion, the margin of error is

$$m = z^* SE$$

- The **confidence level** measures how often the estimation method is successful. A larger confidence level results in a larger margin of error.

Recall the standard error (SE) for population proportions is

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

Some common values for the margin of error:

Confidence Level	Margin of Error
99.7%	$3.0 \cdot SE$
99%	$2.58 \cdot SE$
95%	$1.96 \cdot SE$
90%	$1.645 \cdot SE$
80%	$1.28 \cdot SE$

Example. In 2018, Gallup took a poll of 497 randomly selected adults who teach K–12 students and 42% of them said that digital devices (smartphones, tablets, computers) had "mostly helpful" effects on students' education.

Check that the conditions of the CLT apply.

Estimate the standard error.

Give the 95% confidence interval and interpret the result.

Example. After the Great Recession, the Pew Research Center noted there seemed to be a decline in households that rented their homes and were looking to purchase homes. However, Pew reported that in 2016 “a solid 72%” of renters reported that they wished to buy their own home. Pew reports that the “margin of error at 95% confidence level is plus-or-minus 5.4 points.”

State the confidence interval in interval form and interpret the result.

8.1: The Essential Ingredients of Hypothesis Testing

Definition.

- A **hypothesis test** is a procedure that enables us to choose between two claims.
- The **null hypothesis**, H_0, represents the current belief, or status quo.
- The **alternative hypothesis**, H_a, is what we wish to test.

A hypothesis test has 4 steps:

1. Formulate your null and alternative hypotheses
2. Examine or collect data
3. Compare data to our expectations; is the result significant?
4. Interpret the results

Two-Sided	One-Sided (Left)	One-Sided (Right)
$H_0 : p = p_0$	$H_0 : p = p_0$	$H_0 : p = p_0$
$H_a : p \neq p_0$	$H_a : p < p_0$	$H_a : p > p_0$

Example. When flipping a coin, it is considered fair if both sides of the coin have an equally likely chance of appearing face up. Suppose we have a coin that we believe might be unfair. Let p be the proportion of times where heads appears face up. Formulate the null and alternative hypotheses.

Example. Historically, about 70% of all U.S. adults were married. A sociologist who asks whether marriage rates in the United States have declined will take a random sample of U.S. adults and record whether or not they are married.

Write the null and alternative hypotheses.

Example. An Internet retail business is trying to decide whether to pay a search engine company to upgrade its advertising. In the past 15% of customers who visited the company's web page by clicking on the advertisement bought something. The search engine company offers to do an experiment: for one day a random sample of customers will see the retail business's ad in a more prominent position to try and increase the proportion of customers who make a purchase.

Write the null and alternative hypotheses.

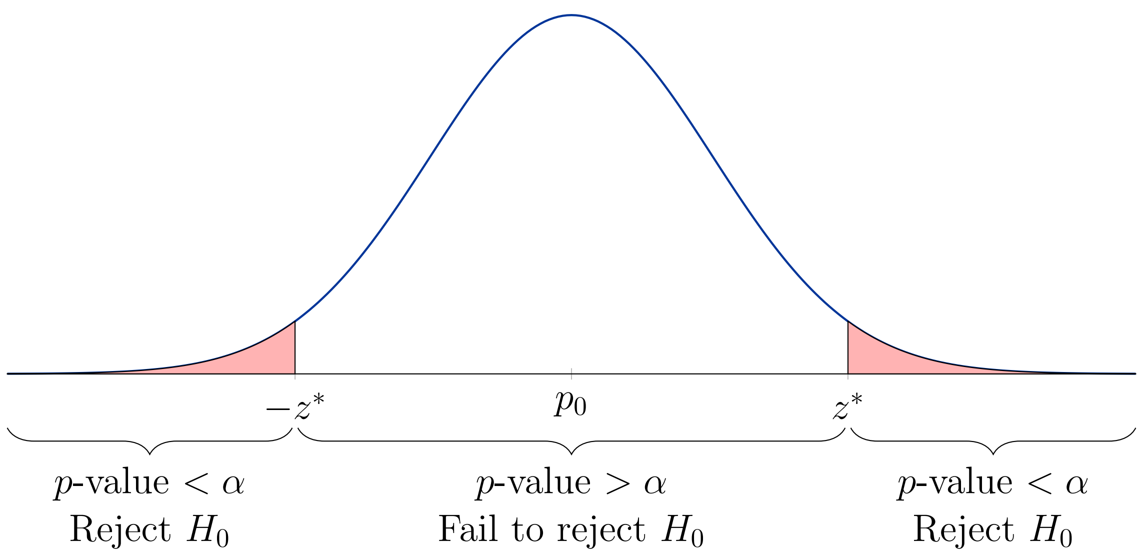
Definition.

- The **significance level**, denoted by α, is the probability of rejecting the null hypothesis when it is actually true (false positive).
- A **test statistic** is similar to a z score comparing the alternative hypothesis to the null hypothesis:

$$z = \frac{\hat{p} - p_0}{SE}, \quad \text{where} \quad SE = \sqrt{\frac{p_0(1-p_0)}{n}}$$

- The p**-value** is the probability that the null hypothesis is true. When the p-value is
  - greater than α, we fail to reject the null hypothesis
  - less than or equal to α, we reject the null hypothesis

Note: Hypothesis tests don't prove the null hypothesis!



8.2: Hypothesis Testing in Four Steps

1. **Hypothesize:** formulate your hypotheses
2. **Check conditions:**
   - **Random and Independent:** The sample must be randomly collected from the population, and observations are independent of each other
   - **Large Sample:** The sample size must be large enough for at least 10 successes, $np_0 \geq 10$, and 10 failures, $n(1-p_0) \geq 10$.
   - **Large Population:** If the sample is collected without replacement, the population of size N must be at least 10 times bigger than the sample: $N \geq 10n$

   If these conditions are met, we compute the test statistic for the One-Proportion z-Test which follows a z-distribution:

   $$z = \frac{\hat{p} - p_0}{SE}, \qquad \text{where} \qquad SE = \sqrt{\frac{p_0(1-p_0)}{n}}$$

3. **Compute:** Stating a significance level, compute the observed test statistic z and/or p-value.
4. **Interpret:** Decide whether to reject or fail to reject the null hypothesis.

Example. Unlike flipping a coin, spinning a coin leads to a biased outcome. Suppose we spun a coin 60 times, and saw a sample proportion of $\hat{p} = 0.35$.

Formulate the null and alternative hypotheses

Check the conditions required to perform a hypothesis test.

Find the test statistic and p-value

Using a significance level of $\alpha = 0.05$, decide whether to reject or fail to reject the null hypothesis.

Example. A group of medical researchers knew from previous studies that in the past, about 39% of all men between the ages of 45 and 59 were regularly active. Researchers were concerned that this percentage had declined over time. For this reason, they did selected a random sample, without replacement, of 1927 men in this age group and interviewed them. Out of this sample, 680 said they were regularly active.

Formulate the null and alternative hypotheses

Check the conditions required to perform a hypothesis test.

Find the test statistic and p-value

Using a significance level of $\alpha = 0.05$, decide whether to reject or fail to reject the null hypothesis.

8.3: Hypothesis Tests in Detail

Definition. (Type I and type II errors)

- A **type I error** is rejecting the null hypothesis, H_0, when it is actually true.
- A **type II error** is failing to reject the null hypothesis, H_0, when it is false.

The probability of committing a type I error is the level of significance: α

	Null Hypothesis is true	Null Hypothesis is false
Reject null hypothesis	Type I error	True positive
Fail to reject null hypothesis	True negative	Type II error

Example. For the following scenarios, identify the type I and type II errors:

"The Boy Who Cried Wolf"

In a court of law, a person is considered innocent until proven guilty.

Testing someone for a disease (e.g. Covid)

Pregnancy test

- When we "fail to reject H_0", we are **not proving** the null hypothesis
- Don't change your hypothesis after you gather your results
- Statistically significant means something likely did not occur by chance
- Confidence intervals vs. Hypothesis testing

Confidence Intervals	Hypothesis Tests
Estimates parameters	Test parameters
Range of values	Is data consistent?

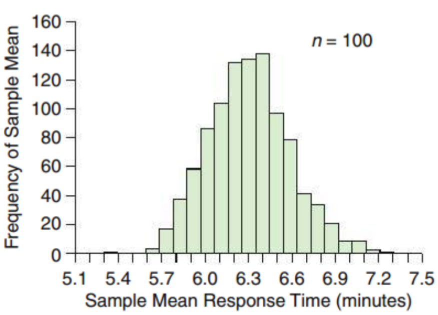
9.1: Sample Means of Random Samples

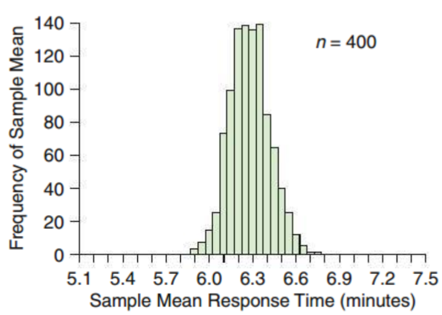
	Sample Statistics	Population Parameters
mean	$\overline{x}$	μ
standard deviation	s	σ
proportion	$\hat{p}$	p

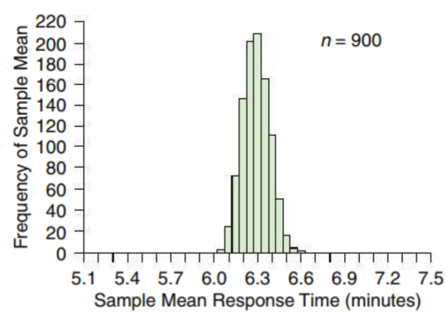
Definition.

- The **sampling distribution** is the distribution of the sample means $\bar{x}$.
- The mean of the sampling distribution is μ so the statistic $\bar{x}$ is an **unbiased estimator**.
- The standard deviation of the sampling distribution is the **standard error**:

$$SE = \frac{\sigma}{\sqrt{n}}$$







9.2: The Central Limit Theorem for Sample Means

Definition. (Central Limit Theorem (CLT))

When estimating a population mean, μ, if

1. *Random and Independent*: Each observation is collected randomly from the population, and observations are independent of each other.
2. *Large Sample*: Either the population distribution is Normal, or the sample size is large ($n \geq 25$).
3. *Big population*: If the sample is collected without replacement (e.g. SRS), then the population size must be at least 10 times bigger than the sample size.

$$N \geq 10n$$

then the sampling distribution for $\overline{x}$ is approximately Normal, with mean μ and standard deviation

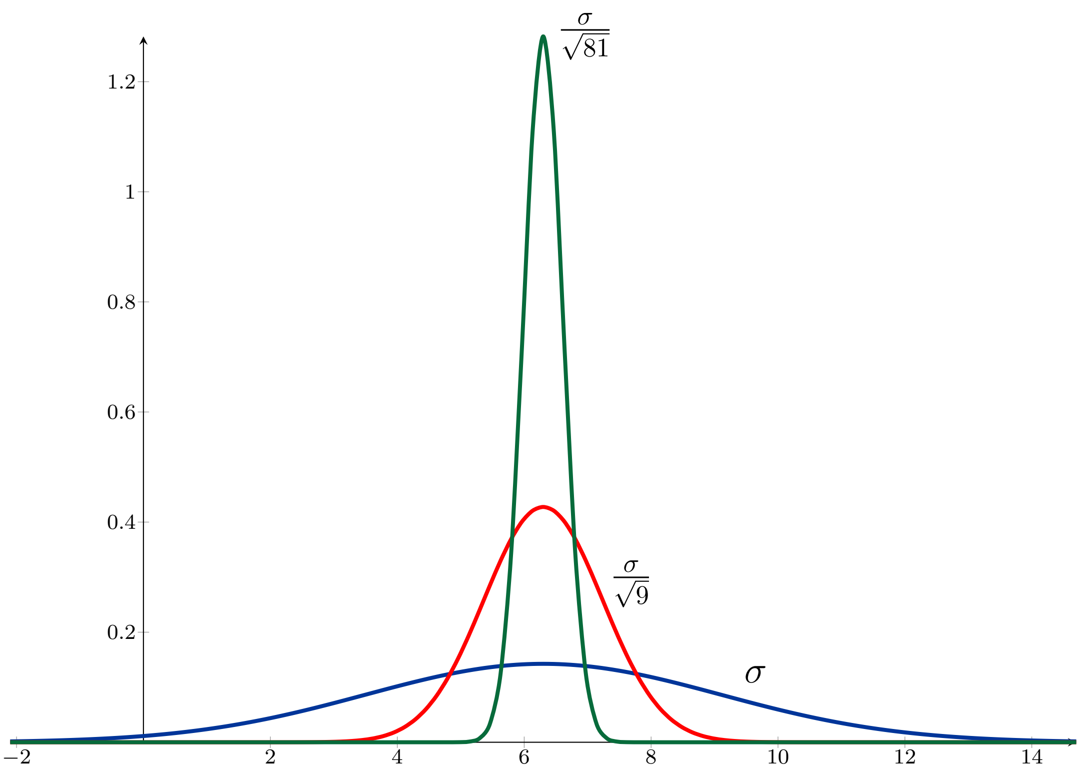
$$SE = \frac{\sigma}{\sqrt{n}}.$$

This distribution is denoted as

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Example. The population distribution of *all* emergency response times from the LA Fire Department is right-skewed. Suppose we repeatedly take random samples of a certain size from this population and calculate the mean response time. We know that the population has mean $\mu = 6.3$ and standard deviation $\sigma = 2.8$ minutes.

Describe the sampling distribution if the sample size is $n = 9$, and again when $n = 81$.



Note: Even if the population distribution has an unusual shape, the sampling distribution is fairly symmetric and unimodal.

Example. According to one very large study done in the US, the mean resting pulse rate of adult women is about $\mu = 74$ BPM, with standard deviation $\sigma = 13$ BPM, where the distribution is known to be skewed right. Suppose we take a random sample of 36 women from this population.

What is the approximate probability that the average pulse rate of this sample will be below 71 or above 77?

Can we find the probability that a single adult woman, randomly selected from this population, will have a resting pulse rate more than 3 BPM away from the mean value, $\mu = 74$?

Definition. (The t-Distribution)
The hypothesis tests and confidence intervals we will use for estimating and testing the mean are based on the **t-statistic**:

$$t = \frac{\overline{x} - \mu}{SE_{est}}$$

$$SE_{est} = \frac{s}{\sqrt{n}}$$

The t-statistic follows the **t-distribution**. With the t-distribution, we do not need to check conditions for the CLT, but the distribution's shape is dependent on the **degrees of freedom (df)**.

If we know the population standard deviation σ, then we have the familiar z-statistic:

$$z = \frac{\overline{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

9.3: Answering Questions about the Mean of a Population

Definition.
Suppose that we wish to estimate a population mean μ based on a sample mean $\overline{x}$.

- A **confidence interval** is an interval about the point estimate $\overline{x}$ that we can be confident contains the true population mean μ:

$$\overline{x} \pm m$$

- The **margin of error (ME)** is half the width of the confidence interval. When estimating a population proportion, the margin of error is

$$m = t^* SE_{\text{est}}$$

  where

$$SE_{\text{est}} = \frac{s}{\sqrt{n}}$$

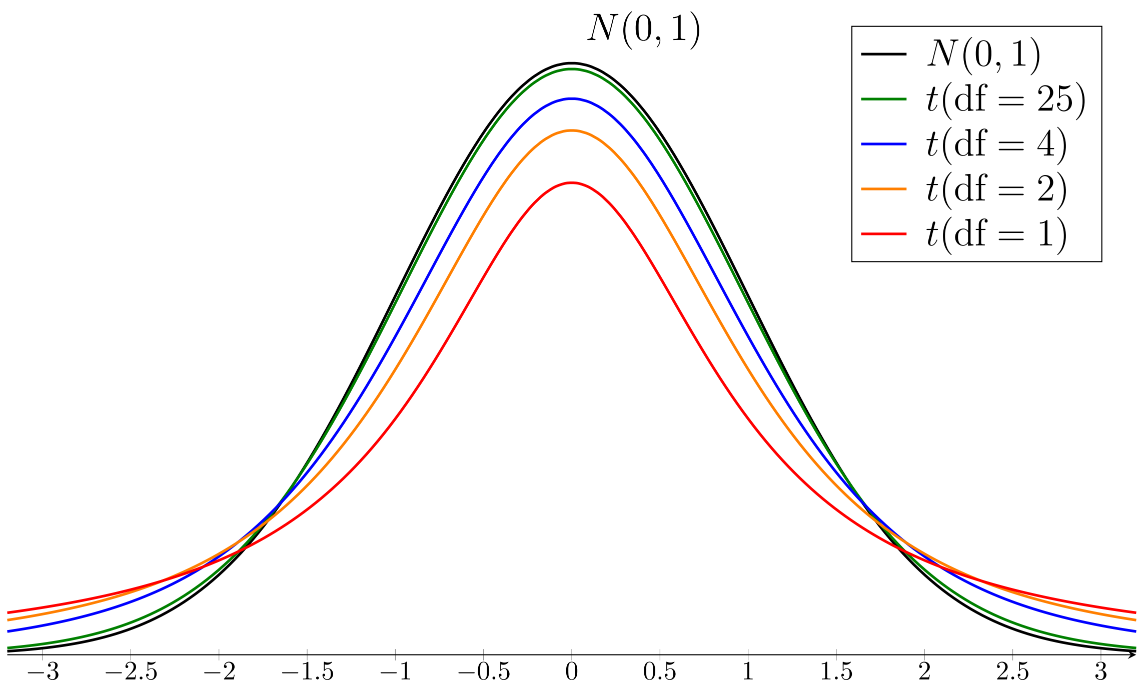
- The **confidence level** measures how often the estimation method is successful. A larger confidence level results in a larger margin of error.

The multiplier t^* is found using the t-distribution and $n - 1$ degrees of freedom.

The t-distribution is used when

- the sample size is small
- the population standard deviation is unknown

As the sample size n (and degrees of freedom, $n-1$) increases, the t-distribution gets closer to the Standard Normal Distribution.



	Confidence Level									
df	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.710	31.820	63.660	318.310	636.620
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Example. A random sample of 35 two-year colleges in 2014-2015 had a mean tuition of \$4,173, with a standard deviation of \$2,590. Find the 90% and 95% confidence intervals and interpret their results.

Example. A study to test the life of iPad batteries reports that in a random sample of 30 iPads, the mean battery life was 9.7 hours, and the standard deviation was 1.2 hours.

Compute a 95% confidence interval and interpret the results.

Suppose Apple claims the iPad has a mean battery life of 10 hours. Based on this confidence interval, do we reject or fail to reject this claim?

Example. Suppose I am interested in estimating my mean commute time. Over the course of a week, I record my commute times:

18.96	20.65	17.47	19.38	17.11

Compute a 95% confidence interval for my actual commute times.

9.4: Hypothesis Testing for Means

Similar to hypothesis testing for proportions, we have the following four steps:

1. **Hypothesize:** formulate your hypotheses
2. **Check conditions:**
   - **Random and Independent:** The sample must be randomly collected from the population, and observations are independent of each other.
   - **Large Sample:** Either the population is Normal, *or* the sample size is large ($n \geq 25$).
   - **Large Population:** If the sample is collected without replacement, the population of size N must be at least 10 times bigger than the sample: $N \geq 10n$

   If these conditions are met, we compute the test statistic for the One-Sample t-Test which follows a t-distribution with $n-1$ degrees of freedom:

   $$t = \frac{\overline{x} - \mu_0}{SE_{\text{est}}}, \qquad \text{where} \qquad SE_{\text{est}} = \frac{s}{\sqrt{n}}$$

3. **Compute:** Stating a significance level, compute the observed test statistic t and/or p-value.
4. **Interpret:** Decide whether to reject or fail to reject the null hypothesis.

Two-Sided	One-Sided (Left)	One-Sided (Right)
$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_a : \mu \neq \mu_0$	$H_a : \mu < \mu_0$	$H_a : \mu > \mu_0$

Example. McDonald's advertises that its ice cream cones have a mean weight of 3.2 ounces. To test this, we find the weights of a sample of 5 cones:

$$4.2, 3.6, 3.9, 3.4, 3.3$$

Formulate the null and alternative hypotheses

Check the conditions required to perform a hypothesis test.

Find the test statistic and p-value

Using a significance level of $\alpha = 0.05$, decide whether to reject or fail to reject the null hypothesis.

Example. In the 2011-2012 academic year, the mean cost of attending two-year colleges in the United States was $\$3,831$. Has this increased over time? A random sample of 35 two-year colleges in 2014-2015 had a mean tuition of $\$4,173$, with a standard deviation of $\$2,590$.

Formulate the null and alternative hypotheses

Check the conditions required to perform a hypothesis test.

Find the test statistic and p-value

Using a significance level of $\alpha = 0.05$, decide whether to reject or fail to reject the null hypothesis.

Repeat this hypothesis test with a sample size of $n = 175$. What happens to the standard error when the sample size increases?